Perturbation-Response and Noise Dynamics in Proteins and Representation Learning for Biomolecular Simulations

by

Yasemin Bozkurt Varolgüneş

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Doctor of Philosophy

in

Electrical and Electronics Engineering

KOÇ
UNIVERSITY

May 2020

Perturbation-Response and Noise Dynamics in Proteins and Representation Learning for Biomolecular Simulations

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral dissertation by

Yasemin Bozkurt Varolgüneş

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Prof. Alper Demir (Advisor)

_____

Prof. Alper T. Erdoğan

_____

Assoc. Prof. Mehmet Sayar

_____

Prof. Canan Atılgan

_____

Assist. Prof. Tristan Bereau

Date: _____

This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK)-BİDEB under 2211 graduate scholarship and 2214-A international collaborative research study support programs.

Dedicated to Serhat

ABSTRACT

Perturbation-Response and Noise Dynamics in Proteins and Representation Learning for Biomolecular Simulations
Yasemin Bozkurt Varolgüneş
Doctor of Philosophy in Electrical and Electronics Engineering
May 2020

Molecular Dynamics simulations, the standard tool for analyzing biomolecules, provide detailed and accurate characterizations but at the expense of tremendous computational cost. A variety of more efficient computational methods have been developed in order to enable the understanding of practical systems of interest. This thesis contributes to this body of work by adapting and repurposing tools from electrical circuit analysis for analyzing the perturbation-response and noise dynamics of proteins, and by applying dimensionality reduction techniques from machine learning for identifying and extracting the essential features of biomolecules from large amounts of simulation data.

The interactions of proteins with ligands are determined by their dynamic characteristics as opposed to only static, time-invariant processes. Inspired by a frequency domain analysis technique from electronic circuit design, we propose a novel computational technique that can be used to analyze small scale functional protein motions as well as interactions with ligands directly in the frequency domain. It can be considered as a generalization of previously proposed static perturbation-response methods, where the frequency of the perturbation becomes the key. We show that the frequency of the perturbation may be an important factor in protein dynamics. Furthermore, we introduce several novel frequency dependent metrics in order to characterize response behavior.

Allostery—a phenomenon in which the binding of a ligand induces alterations in the activity of remote functional sites—can be conceptually viewed as *point-to-point telecommunication in a networked communication medium*, where a signal (ligand) arriving at the input (binding site) propagates through the network (interconnected and interacting atoms) to reach the output (remote functional site). The reliable

transmission of the signal to distal points occurs despite all the disturbances (noise) affecting the protein. Based on this point of view, we propose a computational frequency-domain framework to characterize the displacements and the fluctuations in a region within the protein, originating from the ligand excitation at the binding site and noise, respectively. We characterize the displacements in the presence of the ligand, and the fluctuations in its absence. In the former case, the effect of the ligand is modeled as an external dynamic oscillatory force excitation, whereas in the latter, the sole source of fluctuations is the noise arising from the interactions with the surrounding medium that is further shaped by the internal protein network dynamics. We introduce the excitation frequency as a key factor in a *Signal-to-Noise ratio (SNR)* based analysis, where SNR is defined as the ratio of the displacements stemming from only the ligand to the fluctuations due to noise alone. We then employ an information-theoretic (communication) channel capacity analysis that extends the SNR based characterization by providing a route for discovering new allosteric regions.

Extracting insight from the enormous quantity of data generated from molecular simulations requires the identification of a small number of collective variables whose corresponding low-dimensional free-energy landscape retains the essential features of the underlying system. Data-driven techniques provide a systematic route to constructing this landscape, without the need for extensive a priori intuition into the relevant driving forces. In particular, autoencoders are powerful tools for dimensionality reduction, as they naturally force an information bottleneck and, thereby, a low-dimensional embedding of the essential features. While variational autoencoders ensure continuity of the embedding by assuming a unimodal Gaussian prior, this is at odds with the multi-basin free-energy landscapes that typically arise from the identification of meaningful collective variables. In this work, we incorporate this physical intuition into the prior by employing a Gaussian mixture variational autoencoder (GMVAE), which encourages the separation of metastable states within the embedding. The GMVAE performs dimensionality reduction and clustering within a single unified framework, and is capable of identifying the inherent dimensionality of the input data, in terms of the number of Gaussians required to categorize the data. The resulting embeddings also provide representations for constructing Markov state models, highlighting the transferability of the dimensionality reduction from static equilibrium properties to dynamics.

ÖZETÇE

Proteinlerde Pertürbasyon-Tepki ve Gürültü Dinamiği ve Biyomoleküler Simülasyonlarda Temsili Öğrenme
Yasemin Bozkurt Varolgüneş
Elektrik ve Elektronik Mühendisliği, Doktora
Mayıs 2020

Biyomoleküllerin analizinde standart bir araç olan moleküler dinamik simülasyonları, ayrıntılı ve doğru karakterizasyonlar sağlamakla birlikte bu simülasyonların hesaplama maliyeti yüksektir. İlgilenilen sistemleri anlamayı sağlamak için çeşitli daha verimli hesaplama yöntemleri geliştirilmiştir. Bu tez, proteinlerin pertürbasyon-tepki ve gürültü dinamiklerini analiz etmek için elektrik devresi analizinde kullanılan araçları adapte ederek ve yeniden kullanarak ve makine öğrenmesinde kullanılan boyut azaltma teknikleri aracılığı ile yüksek boyuttaki simülasyon verisinin altında yatan esas nitelikleri çıkartmaya imkan sağlayarak bu alandaki çalışmalara katkı sağlamaktadır.

Proteinlerin ligandlarla etkileşimleri, sadece statik, zamanla değişmeyen işlemlerin aksine dinamik özellikleri ile belirlenir. Elektronik devre tasarımında yaygın olarak kullanılan bir frekans analiz tekniğinden esinlenerek, küçük ölçekli fonksiyonel protein hareketlerinin yanı sıra proteinlerin ligandlarla doğrudan etkileşimlerinin analizinde kullanılabilecek ProteinAC (PAC) adını verdiğimiz yeni bir frekans etki alanı hesaplama tekniği öneriyoruz. Bu teknik daha önce önerilen statik pertürbasyon-tepki yöntemlerinin pertürbasyon frekansının kilit rol oynadığı genelleştirilmesi olarak kabul edilebilir. Pertürbasyon frekansının protein dinamiklerinde önemli bir faktör olabileceğini gösteriyoruz. Ayrıca, tepki davranışını karakterize etmek için frekansa bağlı birkaç yeni metrik sunuyoruz.

Alosteri—ligand bağlanması ile uzak fonksiyonel bölgelerin aktivitesinde değişiklikler gözlemlendiği fenomen—kavramsal olarak bir girişteki (bağlanma bölgesi) sinyalin (ligand) yayılarak çıkışa (uzak etkilenen bölge) ulaştığı ağlı bir iletişim ortamında noktadan noktaya telekomünikasyon olarak kavramsallaştırılabilir. Sinyalin uzak bölgelere kadar güvenilir bir şekilde iletimi, proteini etkileyen tüm bozan

etkenlere (gürültü) rağmen gerçekleşir. Bu bakış açısına dayanarak, proteinin bir bölgesinde meydana gelen bağlanma bölgesine ligand uyarımına bağlı yer değiştirmeleri ve gürültüden kaynaklanan dalgalanmaları frekans etki alanında inceleyen hesaplamalı bir çerçeve öneriyoruz. Ligand varlığındaki yer değiştirmeleri ve yokluğundaki dalgalanmaları karakterize ediyoruz. İlk durumda, ligandın etkisi harici bir dinamik salınım kuvveti uyarımı olarak modellenirken, ikincisinde tek dalgalanma kaynağı, çevre ortamı ile etkileşimlerden kaynaklanan ve dahili protein ağı dinamikleri tarafından şekillendirilen gürültüdür. Uyarım frekansını sadece liganddan kaynaklanan yer değiştirmelerin sadece gürültü nedeniyle oluşana oranı olarak tanımladığımız Sinyal-Gürültü oranına (SNR) dayalı bir analizde anahtar bir faktör olarak tanıtıyoruz. Daha sonra yeni allosterik bölgeleri keşfetmek için bir yol sağlayacak SNR tabanlı karakterizasyonu genişleten bir bilgi teorik (iletişim) kanal kapasite analizi kullanıyoruz.

Moleküler simülasyonlardan elde edilen muazzam miktarda veriden içgörü elde etmek, karşılık gelen düşük boyutlu serbest enerji manzaralarında, altta yatan sistemin temel özelliklerini koruyan az sayıda kolektif değişkenin tanımlanmasını gerektirir. Veriye dayalı teknikler, ilgili itici kuvvetlere dair kapsamlı sezgiye ihtiyaç duymadan bu manzarayı oluşturmak için sistematik bir yol sağlar. Özellikle, oto-kodlayıcılar, doğal olarak bir bilgi darboğazını ve dolayısıyla temel özelliklerin düşük boyutlu olarak gömülmesini zorladıkları için boyutsallığın azaltılmasında kullanılan güçlü araçlardır. Varyasyonel oto-kodlayıcılar, öncül olarak tek modlu bir Gauss varsayarak gömülmenin sürekliliğini sağlarken, bu, tipik olarak anlamlı kolektif değişkenlerin tanımlanmasından kaynaklanan çok havzalı serbest enerji manzaraları ile çelişmektedir. Bu çalışmada, gömülme içinde metastabil durumların ayrılmasını teşvik eden bir Gauss karışımı varyasyonel oto-kodlayıcısı (GMVAE) kullanarak bu fiziksel sezgiyi öncüle dahil ediyoruz. GMVAE, tek bir birleşik çerçeve içinde boyutsallık azaltma ve kümeleme işlemlerini gerçekleştirir ve verileri sınıflandırmak için gereken Gauss dağılımı sayısı bakımından girdi verilerinin doğal boyutsallığını belirleme yeteneğine sahiptir. Ortaya çıkan gösterimler ayrıca, boyut indirgemesinin statik denge özelliklerinden dinamiklere aktarılabilirliğini vurgulayarak Markov durum modellerinin oluşturulması için temsiller sağlar.

PUBLICATIONS

Contents of this thesis have previously appeared in the following publications:

1. Varolgunes, Yasemin Bozkurt, and Alper Demir. "ProteinAC: a frequency domain technique for analyzing protein dynamics." *Physical Biology*, 2018.

2. Varolgunes, Yasemin Bozkurt, Joseph F. Rudzinski, and Alper Demir. "Allostery in proteins as point-to-point telecommunication in a network: Frequency decomposed signal-to-noise ratio and channel capacity analysis." *(Under Review)*

3. Varolgunes, Yasemin Bozkurt, Tristan Bereau, and Joseph F. Rudzinski. "Interpretable embeddings from molecular simulations using Gaussian mixture variational autoencoders." *Machine Learning: Science and Technology*, 2020.

ACKNOWLEDGMENTS

This PhD thesis is the output of the effort and support of several people to whom I am extremely grateful. First and foremost, I wish to express my sincere appreciation to my advisor Prof. Alper Demir for his valuable and constructive suggestions during the planning and development of this thesis. His work discipline, useful critiques, and perfectionist attitude provided me with an invaluable skill set that will benefit me for the rest of my career.

I wish to thank the members of my dissertation committee: Prof. Alper Erdoğan, Assoc. Prof. Mehmet Sayar, Prof. Canan Atılgan, and Assist. Prof. Tristan Bereau for generously offering their time to review this document. I would like to thank to Prof. Burak Erman for his vision and introducing us the topic. I am fully thankful to Prof. Kurt Kremer and Assist. Prof. Tristan Bereau, for welcoming me at the MPIP, and making me feel like a part of the group from the very beginning of my visit.

I would like to acknowledge financial support from the Scientific and Technological Research Council of Turkey (TÜBİTAK), Koç University, and MPIP.

I am indebted to Dr. Joseph F. Rudzinski who has always been a friend indeed, standing by me whenever I needed any kind of support. Advice given by him has been a great help in clarifying my research problems. Thanks for the encouragement and patience throughout the writing process of this thesis. Also thanks for patiently answering my questions about the English language. Who could have thought that our first meeting in Istanbul would evolve like this? I am also excited about our future work.

My appreciation extends to my former and current officemates and colleagues, both at the Koç University and at the MPIP, who provided me with a friendly and

inspiring environment to work and have fun. I would especially like to thank the Theory Group members at the MPIP for facilitating my transition to Germany. Vielen herzlichen dank!

Many thanks to all members of Bozkurt and Varolgüneş families, for always supporting me, even if they did not fully understand all of my work. If my thesis is considered as something to be proud of, that does not compare to how proud I am to be a member of these two wonderful families. I especially thank my mom and dad for allowing me fully realize my own capabilities.

Luckily, I have the luxury of being surrounded by lovely people who care about me, both from my hometown Nazilli and from my college METU. Thanks for always being around when I needed you. I would also like to give a special thanks to Deniz, with whom I have shared moments of deep anxiety. Only in retrospect do I realize how relieving our discussions were in maintaining a healthy mental status during graduate life. Thank you for making life fun as well.

And to my husband Serhat...Thank you so much for playing many different roles in my life and being so good at all of them. Thanks for keeping my scientific curiosity alive and supporting me in all my career choices, including the ones that made/will make us move randomly from one place to another. But most importantly, thanks for making me a better version of myself every day. Without your support, I would have stopped these studies a long time ago. I have no doubt that the biggest stroke of luck in my life is you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AC	Alternating current
ANM	Anisotropic network model
BP	Binding pocket
BW	Bandwidth
CG	Coarse graining/grained
CheY	Chemotaxis signaling protein Y
CK	Chapman-Kolmogorov
CLS	Constrained least squares
CV	Collective variable
DC	Direct current
dRMSD	Distance root mean square deviation
ELBO	Evidence lower bound
ENM	Elastic network model
FBP	Ferric-binding protein
FEL	Free-energy landscape
GNM	Gaussian network model
GMVAE	Gaussian mixture variational autoencoder
KNF	Koshland-Nemethy-Filmer
LHS	Left hand side
LRT	Linear response theory
LTI	Linear time-invariant
MD	Molecular dynamics
MSF	Mean square fluctuations
MSM	Markov state model

MWC	Monod-Wyman-Changeux
NMR	Nuclear magnetic resonance
NN	Neural network
PAC	ProteinAC analysis
PCA	Principal component analysis
PCCA	Perron-cluster cluster analysis
PDC	ProteinDC analysis
PDZ3	Postsynaptic density protein-95/discs large/zonula occludens-1
PRS	Perturbation response scanning
RHS	Right hand side
RMSD	Root mean square deviation
SNR	Signal-to-Noise ratio
TICA	Time-lagged independent component analysis
VAE	Variational autoencoder

Chapter 1

INTRODUCTION

Proteins are the most abundant and diverse macromolecules and perform crucial functions in virtually all biological processes. Structure and dynamics analyses of proteins are essential to understand how they achieve such diverse tasks. Experimental methods such as X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy are mainly utilized for characterizing the three-dimensional structure of proteins. Even though the use of experimental methods is not limited to obtaining a static picture, and more advanced techniques such as NMR relaxation measurements and atomic-force microscopy provide time course of structures, they are somewhat laborious, not suitable for high throughput analyses, and limited in terms of time and space resolution. Numerical simulations come into play to fill this gap, hence they complement and guide the experiments.

Molecular dynamics (MD) simulation is a fundamental tool for understanding a wide range of biological processes, including ligand binding, conformational change, and protein folding [18]. It is a physics-based method where the time evolution of positions and velocities of all atoms is obtained by numerically solving Newton's equations of motion iteratively for every time frame in the simulated time interval. At the end of the simulation what is obtained is a trajectory that describes the three-dimensional motion of the system of interest. Therefore, MD is also called as "computational microscope". Despite being a very useful tool, conventional MD simulations have certain limitations [19]. To better mimick environment conditions, water molecules are added into the simulation box. With this addition, the total number of atoms can easily reach to at least tens of thousands. The evaluation of the non-bonded interactions that act between every pair of atoms yields

high computational cost. The other bottleneck is the limit in the step size used in the numerical calculations. To be able to capture the highest frequency of atomic vibrations, the time step should be limited to several femtoseconds, while the relevant timescale to analyze protein motion is at least on the order of nanoseconds. Thus, millions/trillions of iterations are needed to be performed to simulate such motions. Even with special-purpose hardware such as Anton [20], simulation trajectories rarely reach the timescales of interest. Considering that the interest is in statistical characterization, rather than single anectodal observations, the timescale problem is likely to remain regardless of advances in hardware [21].

MD-related challenges especially become more arduous for instance, if one needs to investigate how proteins respond to specific perturbations, e.g., variations in pH and temperature, chemical modifications, site-specific mutations, interactions with other molecules (e.g., ligand binding) [22]. Unraveling the perturbation-response characteristics without performing separate high-cost simulations for every single perturbation under investigation is particularly important for high throughput screening purposes such drug design applications [23]. Moreover, interpreting simulation results due to the huge amount of trajectory data generated remains as an additional challenge. In this thesis, we propose complementary analysis techniques for the perturbation-response analysis and data representation problems.

In science, it is often encountered that methods or approaches already used in a field of study are translated well to other domains and open new insights. Sometimes it is even observed that the same problems are known by different names in distinct research fields and the equivalences are discovered very later. In the first part of this thesis, we aim to repurpose and adapt well-established techniques from electrical circuit analysis and telecommunications systems, with the motivation of developing fresh and efficient methods that can provide new insights towards "engineering biomolecules". In the second part, we extend the current usage of a machine learning technique for the dynamic modeling and analysis of proteins so as to allow extracting insight from the huge amount of data generated for conformational ensemble of biomolecules.

1.1 Perturbation-response analysis via joint synergies from electrical circuit simulation and telecommunication systems

With some nontrivial modifications, largely due to the protein's three-dimensional nature, an electrical/electronic circuit simulator can be adapted to the analysis-based simulation of proteins. This adaptation is based on the viewpoint of a protein as a network of atoms in analogy to the network of circuit elements.

Some analogies can be readily observed when the circuit and MD simulations are considered. For a review of circuit analysis techniques, the reader is referred to [24], while for the technical details of MD simulations interested reader may refer to [25]. Synergies between these seemingly-different research fields were originally conceptualized in draft form in [57], we present it in a fully-fledged form in this thesis.

- Steady-state analysis:
  - DC analysis solves for an operating point under time-invariant voltage/current excitation by setting the time derivatives of time-dependent components to zero. Then, node voltages and branch currents at the operating point are computed. DC analysis is the first step for other types of analyses: serving as an initial condition for the transient analysis, and as an expansion point for the linearization of all nonlinear components (e.g., diodes and transistors).
  - Similar to the DC analysis of circuits operating under time-invariant excitations, static perturbation methods used in MD, which are based on Linear Response Theory [26, 22], allow calculating displacements from the equilibrium configuration in response to constant force excitations.

- Time simulations:
  - Transient analysis of circuits computes how the voltages and currents change over a specified time interval. Prior to transient analysis, the DC operating point is determined, which is then used as an initial state. Next, the circuit's

response is obtained as a function of time by iteratively solving for the next time points.

– The starting configuration used in MD is generally not in an equilibrium state, therefore energy minimization and equilibration steps are run before the production phase, where the time evolution of particles are monitored.

Sinusoidal steady-state (AC) and noise analyses are among the other frequently-used analysis types in circuit simulations. To the best of our knowledge, direct equivalences of these techniques have not been investigated for MD simulations. Therefore, in the first part of this thesis, we aim to translate these techniques to the MD field, and propose potential applications.

- Sinusoidal steady-state analysis:

  – AC analysis extends the steady-state analysis from constant to time-varying sources by computing the equilibrium response of the circuit to a small sinusoidal excitation, which can be described by its amplitude, frequency, and phase. Calculating the response to a sinusoidal source is particularly important in circuit analysis because of the following: (i) most circuits are driven by AC, i.e., sinusoidal, sources due to the advantages in generation and transmission of AC excitation over DC, and more importantly (ii) it can be used to predict the response to non-sinusoidal sources. Due to the small-signal nature of the excitations, linearization of the nonlinear circuit elements holds around the operating point. Linear time-invariant (LTI) systems excited at a single frequency produce a response that also has a single frequency component at the same frequency, with only a change in amplitude and phase. AC analysis is directly performed in frequency domain, without a need of solving a differential equation, hence no iterative process is necessary. This allows easily obtaining *frequency response*, i.e., how amplitude and phase change at a particular single frequency, over a whole range of different frequencies of interest. Frequency response characterization is an integral part of circuit design, especially indispensable for designing amplifiers and filters, which alter the amplitudes and

phase characteristics with respect to frequency. Some frequently-used filter types include low-pass, band-pass, or high-pass filters that allow passing some frequency intervals while blocking the others, or filters that allow passage of only certain frequencies.

  - We believe that in addition to the static-perturbation analyses, which only examine constant excitations, characterizing the frequency response behavior to time-varying perturbations in an efficient and systematic manner can be advantageous for MD analysis as well. Similar to many benefits it provides in circuit simulations, we advocate that distinct frequency response profiles might even be a determining factor for instance in ligand-binding, or other functional responses of proteins such as allostery.

- Noise analysis:

  - Noise analysis is run together with an AC analysis, and calculates the noise contribution of individual noise-generating element in the circuit to the voltage measured at a specified output node. Noise models for each circuit element/device include one or more of the following noise effects: thermal noise, shot noise, flicker (1/f) noise, and burst noise etc. [27].

  - Particles suspended in a fluid experience constant random fluctuations [28]. It may be crucial to quantify and distinguish the displacements resulting from an external perturbation (such as a ligand) from constant random fluctuations.

Signal-to-Noise ratio (SNR), a frequency-dependent quantity, compares the desired signal power received at the output node to the level of noise power. This concept also opens a connection to telecommunication systems' view, which we briefly review below.

Information Theory, established by the seminal work of Shannon [29], is a scientific field of study that examines the quantification of information in the communication context. The information content of a message is quantified as the expected extent of surprise when delivered to a receiver, i.e., expected amount of clarification

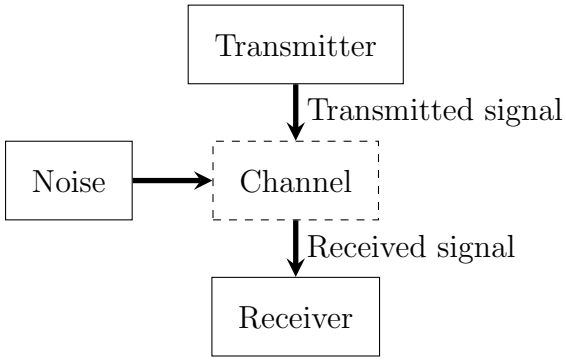


Figure 1.1: Block diagram describing a communication system

of uncertainty. Although much of the early applications center around the classical fields of communication engineering such as telephone lines, fiber-optic communication, radio broadcast, and wireless communication, thanks to Shannon's general framework, it did not take long to be realized that the concepts and methods are also well suited for a broad spectrum of research fields that were not initially thought to be in the scope. Schneider's pioneering study led to the idea that information-theoretical concepts can also be translated to molecular systems—molecular machines, in his terminology [30, 31]. Figure 1.1 schematically presents the common components in any type of communication system: a transmitter, a signal, a channel, and a receiver. The transmitter sends the signal through a medium termed the channel, to the receiver which collects the signal. In physical systems, operation of the channel is subject to noise. We put forward that this general scheme sets a perfect ground for perturbation analysis of biomolecular simulations in the presence of noise, by viewing the intra-protein signal transduction as a multi-channel communication system. In this view, any type of perturbations (be it a ligand, mechanical force, modifications, change in the temperature etc.) is modeled as external force excitation that serves as a transmitter, while the whole protein acts as the medium of the communication, i.e., channel, and the region where the response is probed can be perceived as a receiver. Channel capacity (also known as Shannon limit) is one of the key notions in Shannon's formulation and is defined as the upper bound

for the amount of information the channel can transmit per unit of time, or alternatively the maximum of the mutual information between the input and output of the channel [90]. We propose that the analogy between intra-protein signaling and telecommunication systems becomes very natural, for instance within the allostery context, hence performing similar analyses for biomolecular simulations can be fruitful.

1.2 Representation learning

Extracting insight from the enormous quantity of data generated from molecular simulations requires the identification of a small number of collective variables whose corresponding low-dimensional free-energy landscape retains the essential features of the underlying system. The essential degrees of freedom that define the low-dimensional representation, commonly referred to as collective variables (CVs), are traditionally identified through expert physical/chemical intuition that is often rather specific for the particular system or process of interest [32, 33, 34, 35]. Beyond the characterization of the free-energy landscape (FEL), these CVs can also be used for enhanced sampling [36], or for the construction of low-dimensional configuration-space discretizations, for instance when building Markov state models (MSMs) [37]. The strategy to the discrete-time and discrete-state stochastic models, MSMs, is to utilize multiple shorter trajectories than the timescales of interest. Once the transition can be described in terms of memoryless jumps between the defined states, it also allows describing long-time statistical dynamics, hence addresses the timescale problem. Although the manual selection of CVs can be extremely effective for practitioners with insight into the system, the approach is difficult to extend systematically and is susceptible to missing unanticipated or subtle features of the FEL that may nonetheless play an important role in the relevant phenomena. Data-driven techniques provide an alternative route by inferring the important features directly from the data.

The last couple of years have seen a growing trend towards applying deep neural networks for the automated discovery of collective variables from molecular simula-

tion trajectories [38, 39, 40, 41, 42]. Autoencoders are a natural and powerful tool for dimensionality reduction, as they force an information bottleneck in the latent space. While previous work has demonstrated the potential of these architectures, both for characterizing static properties and for treating kinetics (through the incorporation of a time lag within the autoencoder framework) [43, 38, 44], there has been significantly less work to enforce physical constraints in the latent space. Thus, we propose an architecture that directly incorporates the physical intuition of the multi-basin structure of an ideal free-energy landscape into the latent space prior, representing a significant change in perspective.

1.3 Outline of the thesis

The thesis is composed of two themed parts addressing perturbation-response analysis and representation learning from simulation data, respectively.

Part I begins with **Chapter 2**, where we introduce the basic notation and terminology. The governing equations of motion and related concepts are also introduced.

Chapter 3 lays out mass-spring analogy to derive the force-displacement relationships. We present a generalization for a commonly-utilized scheme, eliminating the requirement to perform the linearization around the equilibrium points.

In **Chapter 4**, we propose a novel computational technique, called ProteinAC (PAC), for analyzing the perturbation-response dynamics of a protein in the frequency domain, which is motivated by alternating current (AC) analysis of electrical networks where the relationship between the magnitude of the voltage/current at the output node is examined through a frequency sweep of a sinusoidal input voltage/current. We show that our method generalizes and subsumes previously proposed static perturbation response methods, by incorporating the frequency of excitation as a key new parameter for dynamic analysis of proteins.

In **Chapter 5**, we propose novel computational techniques for deciphering allosteric phenomena in proteins, that are based on a unique perspective that views the long-range interactions between the binding site (input, transmitter) and the functional site (output, receiver) as point-to-point telecommunication in a networked

communication medium. We propose that, for an allosteric response to occur, the *communication* between the binding site and the functional site should be robust and reliable, despite all the background noise. We make a distinction between the forces originating from the binding event (external force) and the forces that arise from thermal fluctuations (noise) due to solvent interactions. While the external force is directly applied to a local region in the vicinity of the binding pocket, the noise forces act upon every region of the protein. Then, we introduce the perturbation frequency as a key factor in the Signal-to-Noise ratio (SNR) analysis, where the SNR is defined as the ratio of displacements stemming from distinct force types. The information-theoretic channel capacity analysis extends the SNR analysis further and offers new avenues in discovering potentially allosteric regions.

Part II starts with **Chapter 6**, which deals with investigating the use of an extended variational autoencoder framework for dimensionality reduction and clustering of molecular dynamics data. Motivated by the physical intuition that the ideal low-dimensional free-energy landscape will clearly separate the most slowly interconverting metastable states, we adopt a Gaussian mixture model as the prior distribution on the latent space, in contrast to the unimodal Gaussian distribution normally considered. As we demonstrate, this prior counteracts the "anti-clustering" effect often observed in variational autoencoders, leading to clearly separated metastable states in the latent space. We show that the Gaussian mixture variational autoencoder (GMVAE) has significant potential for use as the backbone in a kinetic analysis workflow, as it performs dimensionality reduction and clustering within a single unified framework. Our work validates the methodology on several standard model systems and also a more complex disordered peptide ensemble, and carefully characterizes both the static and kinetic features of the resulting low-dimensional free-energy landscapes.

In **Chapter 7**, a summary is provided with presenting the future research points.

Chapter 8 presents supporting information for the chapters.

Part I

Perturbation-Response Analysis

Chapter 2

BACKGROUND

This chapter provides an overview of protein dynamics starting from the most general form and then narrowing down to a specific linearized state under the paraboloid potential energy surface assumption around the native state. Then, we introduce the transfer function, that quantifies the displacements in response to both external forces and thermal noise, in a frequency decomposed manner. We also provide a short review of normal mode analysis and its use in determining the protein motions.

2.1 Protein dynamics

In principle, the dynamics of a molecular system can be solved quantum mechanically using wave functions that describe the atoms of the system. However, with the large number of atoms in a protein and its surrounding environment, quantum mechanical calculations become impractical for whole protein dynamics. Therefore, a classical mechanics approach is generally used where molecular dynamics describes the time dependent positions and/or velocities of the particles[1].

The dynamics of a protein is governed by its potential energy function, $U(\boldsymbol{r})$, which results from all of the interactions among the particles of the protein. Potential energy of the protein is a function of particle positions, $\boldsymbol{r} = [\boldsymbol{r}_1^T,\ \boldsymbol{r}_2^T,\ \ldots,\ \boldsymbol{r}_N^T]^T$, which is a vector of size $3N \times 1$, where $\boldsymbol{r}_i = [x_i,\ y_i,\ z_i]^T$ is the position of the i^{th}

The content of this chapter has previously appeared in [45].

[1]Here, particle is used as a generic term, independent of the level of detail and the degree of coarse-graining in the model. In all-atom simulations, particles correspond to all of the atoms, whereas in coarse-grained simulations, particles may correspond to a particular set of atoms in the system.

particle ($i = 1, \ldots, N$) and N is the number of particles.

Equations of motion for the i^{th} particle can be written according to Newton's second law as follows

$$\boldsymbol{M}_{(i)} \frac{\mathrm{d}^2}{\mathrm{d}t^2} \boldsymbol{r}_i = -\frac{\partial}{\partial \boldsymbol{r}_i} U(\boldsymbol{r}) + F_{ei}(t) \ , \tag{2.1}$$

where the net force acting upon the particle is composed of an internal force, $f_i(t) = -\frac{\partial}{\partial \boldsymbol{r}_i} U(\boldsymbol{r})$, which is computed as the negative gradient of the potential energy function, and an external force, $F_{ei}(t)$, exerted upon the i^{th} particle, and $\boldsymbol{M}_{(i)} = \mathrm{diag}(m_i, m_i, m_i)$ is a 3×3 diagonal mass matrix with equal diagonal elements corresponding to the mass, m_i, of the i^{th} particle.

When the equations of motions for the particles in a system are combined, the overall dynamics is governed by the following coupled set of $3N$ equations

$$\boldsymbol{M} \frac{\mathrm{d}^2}{\mathrm{d}t^2} \boldsymbol{r} + \frac{\partial}{\partial \boldsymbol{r}} U(\boldsymbol{r}) = \boldsymbol{F}_e(t) \ , \tag{2.2}$$

where $\boldsymbol{M} = \mathrm{diag}(\boldsymbol{M}_{(1)}, \boldsymbol{M}_{(2)}, \ldots, \boldsymbol{M}_{(N)})$ is a block diagonal[2] mass matrix of size $3N$ and $\boldsymbol{F}_e(t) = [\boldsymbol{F}_{e1}(t)^T, \ \boldsymbol{F}_{e2}(t)^T, \ \ldots, \ \boldsymbol{F}_{eN}(t)^T]^T$ is the external force vector of size $3N \times 1$.

The equation of motion, given in Equation 2.2, does not capture the solvent effects due to the interactions with the viscous medium in which the protein resides. In protein modeling, the solvent interactions are either completely ignored (e.g., as in PRS and LRT), or incorporated into the simulations by introducing a large number of solvent particles, or by adding a velocity dependent viscous friction term [46] to the equations of motion approximating the effects of the solvent on the system in a lumped manner as follows

$$\boldsymbol{M} \frac{\mathrm{d}^2}{\mathrm{d}t^2} \boldsymbol{r} + \boldsymbol{L} \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{r} + \frac{\partial}{\partial \boldsymbol{r}} U(\boldsymbol{r}) = \boldsymbol{F}_e(t) \ , \tag{2.3}$$

where $\boldsymbol{L}$ is a diagonal matrix of size $3N$, capturing the effects of the friction or drag caused by the viscosity of the medium the protein is immersed in. This implicit

[2]Since each of the diagonal blocks of $\boldsymbol{M}$, i.e., $\boldsymbol{M}_{(i)}$, is a diagonal matrix corresponding to the i^{th} particle, overall $\boldsymbol{M}$ is a diagonal matrix.

solvent model is preferred since explicit modeling of solvent particles in full-atom MD increases the number of particles in the system significantly. Random effects due to the solvent and other particles can be easily incorporated as an additive term to the RHS, by a random forces $\boldsymbol{\xi}(t)$ to the RHS, leading to the *Langevin* formulation:

$$\boldsymbol{M}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\boldsymbol{r} + \boldsymbol{L}\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{r} + \frac{\partial}{\partial \boldsymbol{r}}U(\boldsymbol{r}) = \boldsymbol{\xi}(t) + \boldsymbol{F_e}(t) \ . \tag{2.4}$$

2.2 Harmonic approximation

In the context of molecular modeling, the *force field* is comprised of the empirical energy functions that include all of the interactions and the parameters needed to calculate the potential energy function of the system. The parameters of the force field are typically determined by fitting to data gathered from experimental studies and utilizing quantum mechanical models for small parts of the system [47]. Generally, most classical force fields are expressed as a summation of terms for various types of interactions within the system such as bond, angle, dihedral, improper terms, and terms for non-bonded interactions. These interaction terms typically contain highly nonlinear functions of particle positions. Hence, potential energy function is almost always non-harmonic. Internal forces, which are the conservative forces imposed by the force field, are defined as the negative of the gradient of the potential energy function with respect to the position vector, $\boldsymbol{r}$. The internal forces computed from the non-harmonic potential energy function are typically highly nonlinear functions of particle displacements, so as the equations of motion. In some methods (e.g., elastic network models (ENMs)), the non-harmonic potential energy function is simplified to a harmonic potential in order to arrive at linear internal force functions, hence linear equations of motion. In other methods, linear(ized) equations are used at intermediate steps. *Linearization* of the force-displacement relationships is performed via a second-order Taylor series expansion of $U(\boldsymbol{r})$ around any given state, i.e., an *expansion point*, $\tilde{\boldsymbol{r}}$, as in the following

$$U(\boldsymbol{r}) \approx U(\tilde{\boldsymbol{r}}) + (\boldsymbol{r}-\tilde{\boldsymbol{r}})^T \frac{\partial}{\partial \boldsymbol{r}}U(\boldsymbol{r})\Big|_{\tilde{\boldsymbol{r}}} + \frac{1}{2}(\boldsymbol{r}-\tilde{\boldsymbol{r}})^T \frac{\partial^2}{\partial \boldsymbol{r}^2}U(\boldsymbol{r})\Big|_{\tilde{\boldsymbol{r}}}(\boldsymbol{r}-\tilde{\boldsymbol{r}}) \ . \tag{2.5}$$

Internal forces, $\boldsymbol{f}(\boldsymbol{r})$, is the negative gradient of the potential energy function $U(\boldsymbol{r})$, and the first-order Taylor series expansion of the internal force functions as shown below corresponds to the second-order expansion above

$$\begin{aligned}\boldsymbol{f}(\boldsymbol{r}) &= -\frac{\partial}{\partial \boldsymbol{r}} U(\boldsymbol{r}) \\ &\approx -\frac{\partial}{\partial \boldsymbol{r}} U(\boldsymbol{r})\Big|_{\tilde{\boldsymbol{r}}} - \frac{\partial^2}{\partial \boldsymbol{r}^2} U(\boldsymbol{r})\Big|_{\tilde{\boldsymbol{r}}} (\boldsymbol{r} - \tilde{\boldsymbol{r}}) \\ &= \boldsymbol{F}_{lin}(\boldsymbol{r})\Big|_{\tilde{\boldsymbol{r}}} ,\end{aligned} \tag{2.6}$$

where $\boldsymbol{F}_{lin}(\boldsymbol{r})$ is a linear function of $\boldsymbol{r}$. The $3N \times 3N$ matrix, $\frac{\partial^2}{\partial \boldsymbol{r}^2} U(\boldsymbol{r})\Big|_{\tilde{\boldsymbol{r}}}$, is the Hessian of the potential energy function evaluated at $\tilde{\boldsymbol{r}}$, i.e., the Jacobian of the internal force function, and it is denoted by $\boldsymbol{H}(\tilde{\boldsymbol{r}})$.

The energy landscape contains many local minima in the conformational space of the protein. According to the folding funnel hypothesis, the native state (equilibrium state) of a protein is the conformation at the free-energy minimum [48]. If the expansion point $\tilde{\boldsymbol{r}}$ is chosen to be the native state of the protein, denoted by $\bar{\boldsymbol{r}}$, then the first order derivatives of the potential energy vanish. This means that the net internal forces acting on the particles of the system at the native state are zero. In this case, the linearized force equation becomes

$$\boldsymbol{F}_{lin}(\boldsymbol{r})\Big|_{\bar{\boldsymbol{r}}} = -\boldsymbol{H}(\bar{\boldsymbol{r}})(\boldsymbol{r} - \bar{\boldsymbol{r}}) \ . \tag{2.7}$$

Hence, Equation 2.4 with the linearized potential energy function reads

$$\boldsymbol{M}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\boldsymbol{\Delta r} + \boldsymbol{L}\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\Delta r} + \boldsymbol{H}(\bar{\boldsymbol{r}})\boldsymbol{\Delta r} = \boldsymbol{\xi}(t) + \boldsymbol{F_e}(t) \ , \tag{2.8}$$

where $\Delta \boldsymbol{r} = \boldsymbol{r} - \bar{\boldsymbol{r}}$ is the deviation/displacement from the native state vector.

2.3 Transfer function

The input-output characteristics for a linear and time-invariant (LTI) system can be described through transfer functions. $\boldsymbol{T}$ represents a matrix of transfer functions from the inputs, the external forces (or the noise) acting on the system represented by $\boldsymbol{F}$, to the output, displacement deviations $\boldsymbol{\Delta r}$

$$\boldsymbol{\Delta r}(f) = \boldsymbol{T}(f) \cdot \boldsymbol{F}(f) \ . \tag{2.9}$$

where f is the frequency of the input and also the output. LTI systems excited at a single frequency f produce a response that also has a single frequency component at the same frequency. Furthermore, LTI systems satisfy the superposition property: The sum of the responses to two different inputs is equal to the response to the sum of the inputs. With $\boldsymbol{F}$ set to either the noise input $\boldsymbol{\xi}(t)$ or the force excitation $\boldsymbol{F_e}(t)$, Equation 2.8 is written in the frequency domain (f) as

$$(\mathrm{i}2\pi f)^2 \boldsymbol{M}\boldsymbol{\Delta r}(f) + \mathrm{i}2\pi f \boldsymbol{L}\boldsymbol{\Delta r}(f) + \boldsymbol{H}(\bar{\boldsymbol{r}})\boldsymbol{\Delta r}(f) = \boldsymbol{F}(f) \ , \tag{2.10}$$

where $\mathrm{i} = \sqrt{-1}$. Thus, complex-valued transfer function $\boldsymbol{T}(f)$ is defined as

$$\begin{aligned} \boldsymbol{T}(f) &= \boldsymbol{K}(f)^{-1} \ , \text{ with} \\ \boldsymbol{K}(f) &= -4\pi^2 f^2 \boldsymbol{M} + \mathrm{i}2\pi f \boldsymbol{L} + \boldsymbol{H}(\bar{\boldsymbol{r}}) \ . \end{aligned} \tag{2.11}$$

2.4 Time-invariant steady state

When random noise forces are not taken into account ($\boldsymbol{\xi} = \boldsymbol{0}$), time derivatives are set to zero, and external force is assumed to be time-invariant, Equation 2.8 simplify to

$$\boldsymbol{H}(\bar{\boldsymbol{r}})\Delta\boldsymbol{r} = \boldsymbol{F}_e \ . \tag{2.12}$$

With rearrangements,

$$\Delta\boldsymbol{r} = \boldsymbol{H}(\bar{\boldsymbol{r}})^{-\boldsymbol{1}}\boldsymbol{F}_e \ , \tag{2.13}$$

which equals to the relationship derived based on thermodynamical formulations from the Linear Response Theory (LRT), which is shown in [26].

Hence, this formulation allows applying an external constant force at a given input particle and calculating the relative displacements of all of the other particles in the protein. In conjunction with a suitable constant force model that mimicks the perturbation of interest, the scheme is useful for calculating the displacements of the perturbed state from the Hessian obtained from the unperturbed state. Within the context of ligand binding, Ikeguchi et al. [26] and Atilgan et al. [22] investigate in detail how to systematically choose the application site upon which the external force acts, as well as the direction and magnitude of the force so as to best mimick

the ligand binding. The input particles that result in displacements that are most similar to the experimentally observed ones between two different forms of a protein (e.g., apo and holo), and the regions with coherent responses within the protein, can be identified.

2.5 Normal mode analysis (NMA)

NMA is a useful technique in trying to elucidate the intrinsic protein dynamics around an energy minimum [49]. Ignoring the solvent effects ($\boldsymbol{L} = \boldsymbol{0}$) and setting the right-hand-side (RHS) of Equation 2.8 to zero yields:

$$\boldsymbol{M}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Delta\boldsymbol{r} = -\boldsymbol{H}(\bar{\boldsymbol{r}})\Delta\boldsymbol{r} \ . \tag{2.14}$$

The general solution to the equation of motion above can be written as a superposition of the normal modes as follows

$$\begin{aligned} \Delta\boldsymbol{r}(t) &= \sum_{i=1}^{3N} \boldsymbol{V}_{hi} \cdot \boldsymbol{\eta}_i(t) \ , \\ \boldsymbol{\eta}_i(t) &= \boldsymbol{C}_i \cdot \cos(\omega_i t + \delta_i) \ . \end{aligned} \tag{2.15}$$

where $\boldsymbol{\eta}_i$ captures the collective motion in the i^{th} normal mode direction. Phase factor, δ_i, and the magnitude $\boldsymbol{C}_i$ are determined by the initial conditions. ω_i is the square root of the eigenvalue and $\boldsymbol{V}_{hi}$ is the corresponding eigenvector of the mass-weighted Hessian, $\boldsymbol{H}_m = (\boldsymbol{M}^{-1}\boldsymbol{H})$. In NMA, the total protein motion is essentially decomposed into a set of independent harmonic oscillators vibrating around their equilibrium state. NMA is based on the eigenvalue decomposition of the Hessian matrix. Since it is real and symmetric around the native state with linearized forces, its eigenvalue decomposition results in

$$\boldsymbol{H}_m = \boldsymbol{V}_h\boldsymbol{\Lambda}{\boldsymbol{V}_h}^T, \ \text{ where } \ \boldsymbol{V}_h{\boldsymbol{V}_h}^T = \boldsymbol{I} \ . \tag{2.16}$$

$\boldsymbol{\Lambda}$ is a $3N \times 3N$ diagonal matrix containing the eigenvalues of the mass-weighted Hessian and the columns of $\boldsymbol{V}_h$ contain the corresponding eigenvectors. In NMA, eigenvectors, $\boldsymbol{V}_{hi}$, are called the normal mode vectors and the eigenvalues, λ_i, are equal to the square of angular normal mode frequencies, ω_i, i.e., $\boldsymbol{\Lambda} = \mathrm{diag}({\omega_1}^2 \ \ldots \ {\omega_{3N}}^2)$.

There are $3N-6$ intrinsic vibrational modes of the system, among the $3N$ solutions, 3 correspond to translations and 3 to rotations of the system.

Low frequency normal modes describe global motions, whereas high frequency modes are for localized movements. The harmonicity assumption leads to a limitation that NMA is valid only around the native state. Still, NMA is used in various applications characterizing the functionally relevant global/local motions, static/mobile regions, rigid/flexible hinge regions, with good fits to experimental data [50].

Chapter 3

APPROXIMATING THE INTERNAL FORCES

In this chapter, we present several techniques that is used in approximating the internal forces, i.e., the negative gradient of the potential energy function. In Section 3.1, we propose a general scheme that allows obtaining the linearized the force-displacement relationships based on mechanical mass-spring analogy around any linearization point. In Section 3.1.1, we show that the presented formulation simplifies when the linearization point is chosen as the equilibrium point. Then, Section 3.2 briefly introduces elastic network models, which represents proteins as an elastic mass-and-spring network, subject to a harmonic potential. These models assume that the linearization point is the equilibrium point, hence the commonly used techniques used in obtaining the linearized internal forces do not generalize for the non-equilibrium points. First, we present two equivalent techniques to achieve linearized forces. Then, we first show that the general scheme we present in Section 3.1 boils down to a reduced-size representation when the linearization is performed around equilibrium points. We emphasize that for the other expansion points, reduced-size models cannot be used, while the technique we present in Section 3.1 maintains its validity. Finally, in Section 3.3, we also show how to retrieve the linearized forces from MD simulations.

3.1 Nodal and node-branch formulations

Force-displacement equations can be written in a systematic manner using a mechanical mass-spring system analogy, where the particles (e.g., atoms, residues) correspond to the joints, i.e., nodes, and the interactions among them are captured

The content of this chapter has previously appeared in [45].

by the springs. If the interactions in the force field among the particles are limited to pairwise interactions only, i.e., if triple and quadruple interactions are removed or expressed in terms of pairwise ones, then any interaction between two particles can be modeled as a nonlinear spring[1]. Thus, a protein can be modeled as a network of connected nonlinear springs. Particles are taken as point masses where the attached springs intersect.

We next describe formulations for computing the steady-state displacements of particles that are connected with nonlinear springs, under a time-invariant external force. Connectivity of the particles and force conservation laws on each particle together with the linearized force on the nonlinear springs are utilized to construct the relationship between the external force and the particle displacements.

When the steady-state condition is assumed, i.e., all the time derivatives are set to zero in Equation 2.2, the forces exerted by the spring on the interacting a^{th} and b^{th} particles, denoted by $\Delta \boldsymbol{f}_{s_{ab}}$[2], depends on the positional distance $\boldsymbol{u}_{ab} = \boldsymbol{r}_a - \boldsymbol{r}_b$, and it is given by

$$\Delta \boldsymbol{f}_{s_{ab}} = \boldsymbol{g}_{ab}(\boldsymbol{u}) = -\frac{\partial}{\partial \boldsymbol{u}_{ab}} U(\boldsymbol{r}) \ , \tag{3.1}$$

where $\boldsymbol{g}_{ab}(\cdot)$ determines the force-displacement relationship. At each point mass, the vectorial sum of the directed forces is set to zero, assuming a static steady-state condition with a time-invariant force excitation $\boldsymbol{F}_e(t) = \boldsymbol{F}_e$. These equations, constituting a *force equilibrium condition*, can be written as follows

$$\boldsymbol{A}_s \Delta \boldsymbol{f}_s = \boldsymbol{F}_e \ . \tag{3.2}$$

$\boldsymbol{A}_s$ is the adjacency matrix (also known as Kirchhoff, connectivity, or incidence matrix) indicating the pairwise interactions among the particles, with entries equal to 0, 1 or -1. Masses attached to the opposite ends of a spring experience a force with

[1]This formulation can be easily extended to the case of triple and quadruple interactions through the conceptualization of a *multi-terminal* spring, one that simultaneously acts on more than two particles.

[2]i^{th} spring is the spring between particles a and b and the force on this spring is denoted by $\Delta \boldsymbol{f}_{s_{ab}} = \Delta \boldsymbol{f}_{s_i}$. Both notations are used interchangeably.

the same magnitude but with opposite directions. An arbitrary direction (aligned with the spring) is assigned for the force exerted by each spring. Then, a force equilibrium equation is written at each point mass. The force of a spring appears with opposite signs in the two equations written at each end. The size of $\boldsymbol{A}_s$ is $3N \times 3N_I$, where N_I is the total number of pairwise interactions (springs) in the system, taking values up to $N^2 - N$ if it is a fully connected system. Generally, the interactions are limited via a cut-off distance r_c. No interaction is assumed if the distance between two particles is larger than r_c. The size of the spring force vector $\Delta \boldsymbol{f}_s$ is $3N_I \times 1$, and $\boldsymbol{F}_e$ is the external force vector of size $3N \times 1$, which represents forces directly applied to the selected particles, e.g., as an excitation or a perturbation.

Similarly, ${\boldsymbol{A}_s}^T$ relates joint positions $\boldsymbol{r}$ to the positional displacements between the two ends of the springs, denoted by $\boldsymbol{u}$, of size $3N_I \times 1$

$$ {\boldsymbol{A}_s}^T \boldsymbol{r} = \boldsymbol{u} \ . \tag{3.3} $$

Spring forces are determined by the nonlinear functions $\boldsymbol{g}_{ab}(\cdot)$. We can simplify the model by linearizing the spring forces around an expansion point $\tilde{\boldsymbol{r}}$, via $\boldsymbol{\alpha}$ given below, a $3N_I \times 3N_I$ Jacobian matrix of the spring forces with respect to $\boldsymbol{u}$. The entries of the Jacobian matrix $\boldsymbol{\alpha}$ can be considered as the spring constants of linear Hookean springs

$$ \Delta \boldsymbol{f}_{slin} = \Delta \boldsymbol{f}_s(\tilde{\boldsymbol{u}}) + \boldsymbol{\alpha} \cdot (\boldsymbol{u} - \tilde{\boldsymbol{u}}) \ \text{ where } \ \boldsymbol{\alpha} = \left. \frac{\partial \Delta \boldsymbol{f}_s}{\partial \Delta \boldsymbol{u}} \right|_{\tilde{\boldsymbol{u}}} , \tag{3.4} $$

$\boldsymbol{\alpha}(\tilde{\boldsymbol{u}}) = \text{diag}(\boldsymbol{\alpha}_{(1)}, \boldsymbol{\alpha}_{(2)}, \ldots, \boldsymbol{\alpha}_{(N_I)})$ is block diagonal and

$$ \boldsymbol{\alpha}_{(i)} = \begin{bmatrix} \dfrac{\partial \Delta \boldsymbol{f}_{si,x}}{\partial \Delta \boldsymbol{u}_{i,x}} & \dfrac{\partial \Delta \boldsymbol{f}_{si,x}}{\partial \Delta \boldsymbol{u}_{i,y}} & \dfrac{\partial \Delta \boldsymbol{f}_{si,x}}{\partial \Delta \boldsymbol{u}_{i,z}} \\ \dfrac{\partial \Delta \boldsymbol{f}_{si,y}}{\partial \Delta \boldsymbol{u}_{i,x}} & \dfrac{\partial \Delta \boldsymbol{f}_{si,y}}{\partial \Delta \boldsymbol{u}_{i,y}} & \dfrac{\partial \Delta \boldsymbol{f}_{si,y}}{\partial \Delta \boldsymbol{u}_{i,z}} \\ \dfrac{\partial \Delta \boldsymbol{f}_{si,z}}{\partial \Delta \boldsymbol{u}_{i,x}} & \dfrac{\partial \Delta \boldsymbol{f}_{si,z}}{\partial \Delta \boldsymbol{u}_{i,y}} & \dfrac{\partial \Delta \boldsymbol{f}_{si,z}}{\partial \Delta \boldsymbol{u}_{i,z}} \end{bmatrix} \tag{3.5} $$

is a 3×3 matrix that is evaluated at $\tilde{\boldsymbol{u}}$. Thus

$$ \Delta \boldsymbol{f}_{slin} - \boldsymbol{\alpha} \Delta \boldsymbol{u} = \Delta \boldsymbol{f}_s(\tilde{\boldsymbol{u}}) \ . \tag{3.6} $$

Equation 3.3 can be expressed in terms of the deviation vectors, $\Delta \boldsymbol{r}$ and $\Delta \boldsymbol{u}$, and can be combined with Equations 3.2 and 3.6 in the form below

$$\begin{bmatrix} \boldsymbol{I} & -\boldsymbol{\alpha}\boldsymbol{A}_s^T \\ \boldsymbol{A}_s & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \Delta \boldsymbol{f}_{slin} \\ \Delta \boldsymbol{r} \end{bmatrix} = \begin{bmatrix} \Delta \boldsymbol{f}_s(\tilde{\boldsymbol{u}}) \\ \boldsymbol{F}_e \end{bmatrix} \ . \tag{3.7}$$

This is the called the *node-branch* or *sparse tableau* formulation [51].

3.1.1 Special case: linearization around equilibrium point

As a special case, the linearization around the native state, $\tilde{\boldsymbol{u}} = \bar{\boldsymbol{u}}$, with $\Delta \boldsymbol{u} = \boldsymbol{u} - \bar{\boldsymbol{u}}$, where normally all of the springs are at their rest lengths with zero force, results in

$$\Delta \boldsymbol{f}_{slin} - \boldsymbol{\alpha} \Delta \boldsymbol{u} = 0 \ , \tag{3.8}$$

yielding

$$\begin{bmatrix} \boldsymbol{I} & -\boldsymbol{\alpha}\boldsymbol{A}_s^T \\ \boldsymbol{A}_s & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \Delta \boldsymbol{f}_{slin} \\ \Delta \boldsymbol{r} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{F}_e \end{bmatrix} \ . \tag{3.9}$$

One can eliminate $\Delta \boldsymbol{f}_{slin}$ in the equations above to obtain

$$(\boldsymbol{A}_s \boldsymbol{\alpha} {\boldsymbol{A}_s}^T) \Delta \boldsymbol{r} = \boldsymbol{F}_e \ . \tag{3.10}$$

The above is called the *nodal* formulation [51]. With this more compact formulation, one can compute the position deviations under external force excitations without first explicitly computing the spring forces, simply as follows

$$\Delta \boldsymbol{r} = \left(\boldsymbol{A}_s \boldsymbol{\alpha} {\boldsymbol{A}_s}^T \right)^{-1} \boldsymbol{F}_e \ . \tag{3.11}$$

At time-invariant steady state (equilibrium), the particle positions do not vary with time. If the term in Equation 2.12 with the time derivative is set to zero, and the resulting equation is compared to Equation 3.10, it can be deduced that $\boldsymbol{A}_s \boldsymbol{\alpha} {\boldsymbol{A}_s}^T$ is equal to the Hessian, $\boldsymbol{H}$, of the potential energy function of the system resulting from the totality of the pairwise interactions. The Hessian $\boldsymbol{H}$ is normally a singular matrix, with six zero eigenvalues, due to the rotational and translational invariance of the system. Therefore, one has to either impose constraints that result in a unique solution, or use an appropriate scheme in solving the nodal equations, e.g., the pseudo-inverse of the Hessian.

3.2 Elastic network models (ENMs)

In ENMs, the protein is represented as a mass-spring network, resulting in a simplified representation of the potential energy function. The nodes of the network, typically chosen as the C_α atoms of the residues, are interconnected by springs. The nominal node positions are determined based on a reference structure, which is usually experimentally determined. The interactions are considered for the residue pairs within a pre-defined cut-off distance of each other. The reference structure is assumed to correspond to the global minimum of the potential energy function by construction, hence no energy minimization is needed. Gaussian Network Models (GNM) [52] and Anisotropic Network Models (ANM) [53] are two subclasses of ENMs. GNMs take into account only the magnitude of the fluctuations, while ANMs capture the directionality as well. ENMs have been successful in predicting the residue fluctuation profiles of globular proteins [54].

Since the linearization point is chosen as the energy minimum in ANMs, construction of Hessian suffices to linearize the internal forces. In the following subsections, two equivalent ways to construct the Hessian are presented: from the second derivative of harmonic potentials, and from the force balance. The equivalence of these approaches is shown in [53].

3.2.1 Hessian from the second order derivative of harmonic potentials

The potential energy function for ANMs can be written as

$$U = \frac{1}{2}\sum_{i<j} \gamma_{ij} |\boldsymbol{\Delta r_i} - \boldsymbol{\Delta r_j}|^2 \ , \tag{3.12}$$

where γ_{ij} is the spring constant between residues i and j, and $|\cdot|$ denotes the L_2 norm (Euclidean length). The Hessian of the potential energy is constructed as

$$\boldsymbol{H_{(ij)}} = -\frac{\gamma_{ij}}{|\boldsymbol{\Delta r_i} - \boldsymbol{\Delta r_j}|^2} \begin{bmatrix} \boldsymbol{r_{j,x}} - \boldsymbol{r_{i,x}} \\ \boldsymbol{r_{j,y}} - \boldsymbol{r_{i,y}} \\ \boldsymbol{r_{j,z}} - \boldsymbol{r_{i,z}} \end{bmatrix} \begin{bmatrix} \boldsymbol{r_{j,x}} - \boldsymbol{r_{i,x}} & \boldsymbol{r_{j,y}} - \boldsymbol{r_{i,y}} & \boldsymbol{r_{j,z}} - \boldsymbol{r_{i,z}} \end{bmatrix} \tag{3.13}$$

for $i \neq j$, where $(\cdot)_{x,y,z}$ denotes the Cartesian coordinates along the corresponding coordinate axes, and the diagonal block for $i = j$ is set to

$$\boldsymbol{H_{(ii)}} = -\sum_{j;j\neq i} \boldsymbol{H_{(ij)}} \ . \tag{3.14}$$

3.2.2 Hessian from force balance

Another commonly-used way of obtaining Hessian in ENMs is based on force-displacement linearization technique discussed in Section 3.1, though utilized in a slightly different manner. In the following, we show that the approach is equivalent to the reduced-size formulation derived for the special equilibrium case that is introduced in Section 3.1.1.

In their reduced-size formulation, instead of examining x, y, z components of the vectors separately, the authors first examine their magnitude and then decompose into components by using directional cosines [22]. This results in reduced size matrices and vectors. The size of $\boldsymbol{\alpha}$, the spring constant matrix, reduces to $N_I \times N_I$ from $3N_I \times 3N_I$, represented by the diagonal $\boldsymbol{\alpha}'$[3]. In ENMs, all entries of $\boldsymbol{\alpha}'$ are usually set to be identical, but they can actually be computed via the force linearization procedure discussed in Section 3.1.

The coordinate components of the spring forces can be related to the magnitude of the spring forces, $|\Delta \boldsymbol{f}_{s(i)}| = \Delta \boldsymbol{f}'_{si}$, as follows

$$\Delta \boldsymbol{f}_{si,x} = \Delta \boldsymbol{f}'_{si} \cdot \cos\theta_i^x \ . \tag{3.15}$$

y and z components are similar. Then, in matrix form for all of the interactions, we obtain the following relationship

$$\Delta \boldsymbol{f}_s = \boldsymbol{C}_s \cdot \Delta \boldsymbol{f}'_s \ , \tag{3.16}$$

where $\boldsymbol{C}_s = \mathrm{diag}(\boldsymbol{C}_{s(1)}, \boldsymbol{C}_{s(2)}, \ldots, \boldsymbol{C}_{s(N_I)})$, $\boldsymbol{C}_{s(i)} = \left[\cos\theta_i^x,\ \cos\theta_i^y,\ \cos\theta_i^z\right]^T$, and θ_i^x, θ_i^y, θ_i^z are the angles between the spring direction vector and the three coordinate axes. $\Delta \boldsymbol{f}'_s$, a vector of size N_I, contains the magnitude of the spring forces.

[3]Prime notation is used for the quantities that are the reduced size equivalents of the formulation described in Section 3.1.

When particles a and b are assumed to be located at the two ends of the spring i, and the coordinates of particle a and b are $\bar{\boldsymbol{r}}_{(a)} = [\bar{x}_a,\ \bar{y}_a,\ \bar{z}_a]^T$, $\bar{\boldsymbol{r}}_{(b)} = [\bar{x}_b,\ \bar{y}_b,\ \bar{z}_b]^T$, then the angles are calculated as follows

$$\boldsymbol{C}_{s(i)} = \begin{bmatrix} \cos\theta_i^x \\ \cos\theta_i^y \\ \cos\theta_i^z \end{bmatrix} = \begin{bmatrix} \bar{x}_a - \bar{x}_b \\ \bar{y}_a - \bar{y}_b \\ \bar{z}_a - \bar{z}_b \end{bmatrix} / |\bar{\boldsymbol{r}}_a - \bar{\boldsymbol{r}}_b| \ . \tag{3.17}$$

The reduced size vectors and matrices used by ENMs can be obtained with the following when the protein is assumed to be in the native state where the net internal forces acting on the particles are zero. For differently chosen linearization points, the reduced size formulation in ENMs is not equivalent to the general formulation described in Section 3.1.

$$\begin{aligned} \boldsymbol{A}'_s &= \boldsymbol{A}_s \cdot \boldsymbol{C}_s \\ \Delta\boldsymbol{u} &= \boldsymbol{C}_s \cdot \Delta\boldsymbol{u}' \\ \boldsymbol{\alpha} &= \boldsymbol{C}_s \cdot \boldsymbol{\alpha}' \cdot \boldsymbol{C}_s^T \\ \Delta\boldsymbol{f}_s &= \boldsymbol{C}_s \cdot \Delta\boldsymbol{f}'_s \\ \Delta\boldsymbol{r} &= \Delta\boldsymbol{r}' \\ \boldsymbol{F}_e &= \boldsymbol{F}'_e \ . \end{aligned} \tag{3.18}$$

Here, we will derive the relationship between $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$, the others are similar. For spring i, we have the following Jacobian matrix that is obtained with the scheme in Section 3.1 and $\boldsymbol{\alpha}'_i$ is defined as follows

$$\boldsymbol{\alpha}_{(i)} = \frac{\partial \Delta\boldsymbol{f}_{s(i)}}{\partial \Delta\boldsymbol{u}_{(i)}}, \ \ \boldsymbol{\alpha}'_i = \frac{\partial \Delta\boldsymbol{f}'_{si}}{\partial \Delta\boldsymbol{u}'_i} \ . \tag{3.19}$$

$\Delta\boldsymbol{f}_s$ for the i^{th} spring in Equation 3.16, is inserted into the $\boldsymbol{\alpha}_{(i)}$ equation above.

$$\boldsymbol{\alpha}_{(i)} = \frac{\partial \Delta\boldsymbol{f}_{s(i)}}{\partial \Delta\boldsymbol{u}_{(i)}} = \frac{\partial (\boldsymbol{C}_{s(i)} \Delta\boldsymbol{f}'_{si})}{\partial \Delta\boldsymbol{u}_{(i)}} \ . \tag{3.20}$$

Then, by using the chain rule for differentiation, the following is obtained

$$\boldsymbol{\alpha}_{(i)} = \frac{\partial \boldsymbol{C}_{s(i)}}{\partial \Delta\boldsymbol{u}_{(i)}} \Delta\boldsymbol{f}'_{si} + \boldsymbol{C}_{s(i)} \frac{\partial \Delta\boldsymbol{f}'_{si}}{\partial \Delta\boldsymbol{u}_{(i)}} \ . \tag{3.21}$$

Partial derivative in the second term in the summation above is written as a multiplication of partial derivatives by using the chain rule and the following relationship: $\boldsymbol{C}_{s(i)} = \dfrac{\Delta \boldsymbol{u}_{(i)}}{\Delta \boldsymbol{u}'_i}$

$$\frac{\partial \Delta \boldsymbol{f}'_{si}}{\partial \Delta \boldsymbol{u}_{(i)}} = \frac{\partial \Delta \boldsymbol{f}'_{si}}{\partial \Delta \boldsymbol{u}'_i} \frac{\partial \Delta \boldsymbol{u}'_i}{\partial \Delta \boldsymbol{u}_{(i)}} = \boldsymbol{\alpha}'_i \boldsymbol{C}_{s(i)}^{\,T} \ . \tag{3.22}$$

Then, the second part of Equation 3.21 becomes $\boldsymbol{\alpha}'_i \boldsymbol{C}_{s(i)} \boldsymbol{C}_{s(i)}^{\,T}$ and the first part is zero ($\Delta \boldsymbol{f}'_{si} = 0$) when the native state conditions are assumed. Thus, the following relationship becomes valid: $\boldsymbol{\alpha}_{(i)} = \boldsymbol{\alpha}'_i \boldsymbol{C}_{s(i)} \boldsymbol{C}_{s(i)}^{\,T}$.

However, at points other than the native state, the derivative in the first part of Equation 3.21 becomes

$$\frac{\partial \boldsymbol{C}_{s(i)}}{\partial \Delta \boldsymbol{u}_{(i)}} = \frac{1}{\Delta \boldsymbol{u}'_i} [\boldsymbol{I}_3 - \boldsymbol{C}_{s(i)} \boldsymbol{C}_{s(i)}^{\,T}] \ . \tag{3.23}$$

Then, Equation 3.21 can be written as

$$\boldsymbol{\alpha}'_i = \frac{\Delta \boldsymbol{f}'_{si}}{\Delta \boldsymbol{u}'_i} [\boldsymbol{I}_3 - \boldsymbol{C}_{s(i)} \boldsymbol{C}_{s(i)}^{\,T}] + \boldsymbol{\alpha}'_i \boldsymbol{C}_{s(i)} \boldsymbol{C}_{s(i)}^{\,T} \ . \tag{3.24}$$

When the above is repeated for all the coordinates and interactions, the general relationship between $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ is formulated as follows

$$\boldsymbol{\alpha} = \boldsymbol{K} \otimes \boldsymbol{I}_3 - \boldsymbol{C}_s \boldsymbol{K} \boldsymbol{C}_s^T + \boldsymbol{C}_s \boldsymbol{\alpha}' \boldsymbol{C}_s^{\,T} \ . \tag{3.25}$$

Here, $\otimes$ denotes the Kronecker product operation and
$\boldsymbol{K} = \text{diag}\left(\dfrac{\Delta \boldsymbol{f}'_{s1}}{\Delta \boldsymbol{u}'_1}, \dfrac{\Delta \boldsymbol{f}'_{s2}}{\Delta \boldsymbol{u}'_2}, \ldots, \dfrac{\Delta \boldsymbol{f}'_{sN_I}}{\Delta \boldsymbol{u}'_{N_I}} \right)$.

This shows that the original ENM formulation does not capture the case when the linearization is performed around a non-equilibrium point in the conformational space of the protein. However, the scheme that we propose in Section 3.1 is more general and maintains its validity even for linearization around non-equilibrium points. Note that $\boldsymbol{\alpha}' = \boldsymbol{C}_s^T \boldsymbol{\alpha} \boldsymbol{C}_s$ satisfies Equation 3.25. Note also that $\boldsymbol{C}_s^T \boldsymbol{C}_s = \boldsymbol{I}$ and $\boldsymbol{C}_s \boldsymbol{C}_s^T \neq \boldsymbol{I}$.

Thus, we obtain the proposed relationship around the native state: $\boldsymbol{\alpha} = \boldsymbol{C}_s \boldsymbol{\alpha}' \boldsymbol{C}_s^{\,T}$. Then, the reduced *node-branch* and *nodal* formulations around the native state can be expressed respectively as follows

$$\begin{bmatrix} \boldsymbol{I} & -\boldsymbol{\alpha}' \boldsymbol{A}_s'^T \\ \boldsymbol{A}'_s & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \Delta \boldsymbol{f}'_s \\ \Delta \boldsymbol{r}' \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{F}'_e \end{bmatrix} \ , \tag{3.26}$$

$$(\boldsymbol{A}'_s \boldsymbol{\alpha}' {\boldsymbol{A}_s}'^T) \Delta \boldsymbol{r}' = \boldsymbol{F}'_e \ . \tag{3.27}$$

One may observe that the coefficient matrix $(\boldsymbol{A}'_s \boldsymbol{\alpha}' {\boldsymbol{A}_s}'^T)$ on the LHS of Equation 3.27 is in fact the Hessian matrix $\boldsymbol{H}$, and

$$\Delta \boldsymbol{r} = \boldsymbol{H}^{-1} \boldsymbol{F}_e \ . \tag{3.28}$$

3.3 Hessian from MD

In addition to obtaining force-displacement relationship, here we present how to get hessian from MD. As the first step, positional fluctuation trajectories around a reference or equilibrium structure are calculated. Next, the covariance matrix is formed based on the time averages over the trajectories as

$$\mathbf{Cov}_{ij} = \begin{bmatrix} \langle \Delta \boldsymbol{r}_{i,x} \Delta \boldsymbol{r}_{j,x} \rangle & \langle \Delta \boldsymbol{r}_{i,x} \Delta \boldsymbol{r}_{j,y} \rangle & \langle \Delta \boldsymbol{r}_{i,x} \Delta \boldsymbol{r}_{j,z} \rangle \\ \langle \Delta \boldsymbol{r}_{i,y} \Delta \boldsymbol{r}_{j,x} \rangle & \langle \Delta \boldsymbol{r}_{i,y} \Delta \boldsymbol{r}_{j,y} \rangle & \langle \Delta \boldsymbol{r}_{i,y} \Delta \boldsymbol{r}_{j,z} \rangle \\ \langle \Delta \boldsymbol{r}_{i,z} \Delta \boldsymbol{r}_{j,x} \rangle & \langle \Delta \boldsymbol{r}_{i,z} \Delta \boldsymbol{r}_{j,y} \rangle & \langle \Delta \boldsymbol{r}_{i,z} \Delta \boldsymbol{r}_{j,z} \rangle \end{bmatrix} \ .$$

$\mathbf{Cov}_{ij}$ above is a 3×3 matrix of the covariances between the x, y, z components of the positional deviation vectors $\Delta \boldsymbol{r}_i$ and $\Delta \boldsymbol{r}_j$ for nodes i and j. The overall covariance matrix $\mathbf{Cov}$ is then formed as a block matrix of size $3N \times 3N$, where the ij^{th} block is set to $\mathbf{Cov}_{ij}$. The Hessian matrix is in fact equal to the scaled inverse of the covariance matrix, $\boldsymbol{H}^{\mathrm{MD}} = \mathrm{k_B T}\, \mathbf{Cov}^{-1}$ [55], where $\mathrm{k_B}$ is the Boltzmann's constant and T is the absolute temperature.

Chapter 4

PAC: A FREQUENCY DOMAIN TECHNIQUE FOR ANALYZING PROTEIN DYNAMICS

4.1 Introduction

In static perturbation studies such as Linear Response Theory (LRT) [26] and and Perturbation Response Scanning (PRS) [22], the ligand is modeled via a constant force vector exerted on the binding site residue. As such, the dynamical, time-varying behavior of vibratory ligand-protein interactions can not be investigated. Based on thermodynamical arguments, the ligand binding process is characterized as a dynamic event with associated vibratory and fluctuation-like motions as opposed to permanent binding [56]. Therefore, in addition to the directionality and the amplitude, the *frequency* of the applied force representing a perturbation may be a determining factor in binding/unbinding, and other functional responses of the protein such as allostery.

We propose a novel computational technique, namely *Protein AC Analysis (PAC)*, for analyzing the perturbation-response dynamics of a protein directly in the frequency domain. Motivated by AC analysis of electrical networks, where the circuit is excited by a sinusoidal input at an input node and the amplitude/phase responses are monitored at the output node(s) as a function of frequency, our proposed technique makes it possible to analyze the behavior of a protein under excitations with varying frequencies, which was not possible with previously proposed methods.

PAC is distinct from, but related to, normal mode analysis (NMA) [58] that also produces a sort of frequency domain characterization for protein dynamics. NMA is useful in determining the global and local motions of a protein around an equilibrium

The content of this chapter has previously appeared in [45].

and it is based on intrinsic, autonomous dynamics. The frequency in PAC is instead determined by the excitation on the system as an external factor.

There are other approaches to frequency domain analysis of proteins in the literature. Notably, in [46], the authors describe a frequency decomposed fluctuation analysis formulation for proteins that is compared with the standard formulation based on thermodynamics. Their formulation allows one to compute contributions to the fluctuation dynamics in frequency intervals of interest, which is not possible with the thermodynamics formulation. PAC is similar in spirit to the work described in [46] based on a Langevin equation formulation. The technique described in [46] enables fluctuation analysis in the frequency domain, whereas PAC extends static perturbation-response analysis in order to characterize the frequency response of proteins.

PAC is not limited to the residue based coarse grained models, it can in fact be used in conjunction with all-atom models, possibly following energy minimization in MD. In order to demonstrate the utility of the proposed method, here we describe the PAC technique in conjunction with a coarse-grained anisotropic network model (ANM), a subclass of ENMs, constructed at an energy minimized native state. As we show, our technique is able to capture the response behavior computed by the static perturbation methods and can be regarded as a generalization of these techniques via the introduction of the frequency of the input perturbations that becomes crucial in studying ligand binding dynamics.

The chapter is organized as follows: In Section 4.2, we introduce the proposed PAC technique and describe how it can be used to analyze the displacements of the particles under force excitations with varying frequency and show that PAC is in fact a generalization of the static perturbation methods, with static methods obtained from PAC by setting the frequency of excitation to zero. Furthermore, in Section 4.3, we propose a set of frequency dependent metrics that help elucidate the complex motions of residues/atoms in a compact and informative manner under varying frequency force excitations. We describe a protocol for the application of PAC to the analysis of protein dynamics in Section 4.4. In Section 4.5, we present

results for the ferric binding protein and bacterial chemotaxis protein by computing the proposed metrics via PAC.

4.2 Theory

PAC is based on the linearization of the system dynamics around an expansion point (e.g., an operating point, DC point, conformation, or native state) as in AC analysis of electrical networks. For PAC, the natural choice for the expansion point is the native state of the protein, but others are possible. We consider the linearized force field around the native state (please see Chapter 2 for the details of linearization), and in terms of the deviation from the native state vector, $\Delta\boldsymbol{r} = \boldsymbol{r} - \bar{\boldsymbol{r}}$ as described in Equation 2.8 . Here, we do not take into account the random forces. The variables are as introduced in Chapter 2: $\boldsymbol{M}$ stands for the masses, $\boldsymbol{L}$: friction, $\boldsymbol{H}$: linearized potential energy function, $\boldsymbol{F_e}$: external forces.

$$\boldsymbol{M}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\boldsymbol{\Delta r} + \boldsymbol{L}\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\Delta r} + \boldsymbol{H}(\bar{\boldsymbol{r}})\boldsymbol{\Delta r} = \boldsymbol{F_e}(t)\ , \tag{4.1}$$

Next, we transform Equation 4.1 to state-space form with the state vector $\boldsymbol{v} = [\Delta\boldsymbol{r}^T,\ \Delta\dot{\boldsymbol{r}}^T]^T$, where $\Delta\dot{\boldsymbol{r}} = \frac{\mathrm{d}}{\mathrm{d}t}\Delta\boldsymbol{r}$

$$\begin{bmatrix}\Delta\dot{\boldsymbol{r}}\\ \Delta\ddot{\boldsymbol{r}}\end{bmatrix} = \underbrace{\begin{bmatrix}\boldsymbol{0} & \boldsymbol{I}\\ -\boldsymbol{M}^{-1}\boldsymbol{H} & -\boldsymbol{M}^{-1}\boldsymbol{L}\end{bmatrix}}_{\boldsymbol{A}}\begin{bmatrix}\Delta\boldsymbol{r}\\ \Delta\dot{\boldsymbol{r}}\end{bmatrix} + \underbrace{\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{M}^{-1}\end{bmatrix}}_{\boldsymbol{B}}\boldsymbol{F}_e(t)$$

$$\dot{\boldsymbol{v}} = \boldsymbol{A}\boldsymbol{v} + \boldsymbol{B}\boldsymbol{F}_e(t)\ . \tag{4.2}$$

We introduce the *output observable* vector $\boldsymbol{y}$ of size $3N \times 1$, as linear combinations of the first set of the state variables in $\boldsymbol{v}$, $\Delta\boldsymbol{r}$. $\boldsymbol{C}$, a matrix of size $3N \times 6N$, is constructed in such a way so that the last $3N$ columns are zero, i.e., $\boldsymbol{C} = \begin{bmatrix}\boldsymbol{C}_1^{3N\times 3N} & \boldsymbol{0}^{3N\times 3N}\end{bmatrix}$. In order to probe the displacement of every particle, $\boldsymbol{C}_1$ is set to the identity matrix of size 3N. However, it is also possible to examine the collective movements (e.g., whole protein motion, a domain's motion) with an ap-

propriately constructed $\boldsymbol{C}_1$.

$$\begin{aligned} \dot{\boldsymbol{v}} &= \boldsymbol{A}\boldsymbol{v} + \boldsymbol{B}\boldsymbol{F}_e(t) \\ \boldsymbol{y} &= \boldsymbol{C}\boldsymbol{v} \ . \end{aligned} \tag{4.3}$$

Next, we transform the state space formulation in Equation 4.3 to the Laplace, i.e., frequency, domain to obtain

$$\begin{aligned} s\boldsymbol{V}(s) &= \boldsymbol{A}\boldsymbol{V}(s) + \boldsymbol{B}\boldsymbol{F}_e(s) \\ \boldsymbol{Y}(s) &= \boldsymbol{C}\boldsymbol{V}(s) \end{aligned} \tag{4.4}$$

that turns the differential equation into an algebraic one with the frequency of interest as a parameter. The transfer function (or complex valued gain of the system) is represented by $\boldsymbol{T}$ as follows

$$\begin{aligned} \boldsymbol{Y}(s) &= \boldsymbol{T}(s) \cdot \boldsymbol{F}_e(s) \\ &= \boldsymbol{C}[s\boldsymbol{I} - \boldsymbol{A}]^{-1}\boldsymbol{B}\boldsymbol{F}_e(s) \ , \end{aligned} \tag{4.5}$$

where s is the complex frequency. In PAC, we evaluate s on the imaginary axis, i.e., $s = \mathrm{i}2\pi f$ where $\mathrm{i} = \sqrt{-1}$, and f is the frequency of excitation. The complex valued transfer function, $\boldsymbol{T}$, takes varying values as a function of frequency, therefore, the response characteristics depend not only on the magnitudes, directions or points of application but also the frequency of excitations. $\boldsymbol{T}(s)$, with size $3N \times 3N$, is the matrix of complex valued gains along the coordinate axes, consisting of 3×3 block matrices and constructed in the following form

$$\boldsymbol{T}(s) = \begin{bmatrix} \boldsymbol{T}_{(11)}(s) & \boldsymbol{T}_{(12)}(s) & \dots & \boldsymbol{T}_{(1N)}(s) \\ \boldsymbol{T}_{(21)}(s) & \boldsymbol{T}_{(22)}(s) & \dots & \boldsymbol{T}_{(2N)}(s) \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{T}_{(N1)}(s) & \boldsymbol{T}_{(N2)}(s) & \dots & \boldsymbol{T}_{(NN)}(s) \end{bmatrix} \ . \tag{4.6}$$

Submatrix $\boldsymbol{T}_{(ij)}(s)(i, j = 1, \dots, N)$ of size 3×3 contains the transfer functions between the output i and the external force that is applied on the j^{th} particle. PAC does not pose any limitation on the form of the excitation force function, $\boldsymbol{F}_e$, provided that the Laplace transform exists. In this thesis, however, we consider

force excitations that are sinusoidal in time. We apply a force excitation to one of the particles, j, in the form

$$\boldsymbol{F}_{ej} = [F_{j_x}\cos(\omega t + \phi_{j_x}),\ \ F_{j_y}\cos(\omega t + \phi_{j_y}),\ \ F_{j_z}\cos(\omega t + \phi_{j_z})]^T\ , \tag{4.7}$$

and set all other entries in $\boldsymbol{F}_e(t)$ to zero. F_{j_x}, F_{j_y}, and F_{j_z} are the amplitudes of the force applied on the j^{th} particle, and ϕ_{j_x}, ϕ_{j_y}, and ϕ_{j_z} are the phase shifts in the $x,\ y,\ z$ directions, respectively, and $\omega = 2\pi f$ is the angular frequency. Here, we assume that the amplitudes of the force components are sufficiently small so that the linearized force field is a good approximation. Due to this small displacement assumption, at steady-state, the response/output displacements are also sinusoidal at the same frequency of the force excitation, but with frequency dependent phase shifts and amplitudes. Steady-state dynamic response of the i^{th} output $\boldsymbol{y}^{ss}_{(i)}(t)$, to the sinusoidal force excitation can be computed using the transfer functions, where $\boldsymbol{T}_{(ij)}(\mathrm{i}\omega) = \boldsymbol{T}_{(ij)}(s)\Big|_{s=\mathrm{i}\omega}$ is of the form

$$\boldsymbol{T}_{(ij)}(s) = \begin{bmatrix} t_{xx(ij)}(s) & t_{xy_{(ij)}}(s) & t_{xz(ij)}(s) \\ t_{yx_{(ij)}}(s) & t_{yy_{(ij)}}(s) & t_{yz_{(ij)}}(s) \\ t_{zx(ij)}(s) & t_{zy_{(ij)}}(s) & t_{zz(ij)}(s) \end{bmatrix}\ , \tag{4.8}$$

and each complex entry of $\boldsymbol{T}_{(ij)}$ is written in exponential form as follows

$$t_{xx(ij)}(\mathrm{i}\omega) = |t_{xx(ij)}(\mathrm{i}\omega)|\exp\left(\mathrm{i}\theta_{xx(ij)}(\omega)\right)\ , \tag{4.9}$$

where $|t_{xx(ij)}| = \sqrt{\mathrm{Im}\{t_{xx(ij)}\}^2 + \mathrm{Re}\{t_{xx(ij)}\}^2}$[1] and $\theta_{xx(ij)} = \tan^{-1}(\frac{\mathrm{Im}\{t_{xx(ij)}\}}{\mathrm{Re}\{t_{xx(ij)}\}})$ is the magnitude and the phase of $t_{xx(ij)}$, respectively. Then, the i^{th} steady-state output

[1]Re and Im is the real and imaginary part of a complex number.

$\boldsymbol{y}^{ss}_{(i)}(t)$ is as follows

$$\begin{aligned}\boldsymbol{y}^{ss}_{(i)}(t) &= \begin{bmatrix} y^{ss}_{i,x} & y^{ss}_{i,y} & y^{ss}_{i,z} \end{bmatrix}^T \\ &= \begin{bmatrix} |t_{xx(ij)}(\mathrm{i}\omega)|F_{j_x}\cos(\omega t + \phi_{j_x} + \theta_{xx(ij)}(\omega)) + \\ |t_{xy(ij)}(\mathrm{i}\omega)|F_{j_y}\cos(\omega t + \phi_{j_y} + \theta_{xy(ij)}(\omega)) + \\ |t_{xz(ij)}(\mathrm{i}\omega)|F_{j_z}\cos(\omega t + \phi_{j_z} + \theta_{xz(ij)}(\omega)) \\ \hline |t_{yx(ij)}(\mathrm{i}\omega)|F_{j_x}\cos(\omega t + \phi_{j_x} + \theta_{yx(ij)}(\omega)) + \\ |t_{yy(ij)}(\mathrm{i}\omega)|F_{j_y}\cos(\omega t + \phi_{j_y} + \theta_{yy(ij)}(\omega)) + \\ |t_{yz(ij)}(\mathrm{i}\omega)|F_{j_z}\cos(\omega t + \phi_{j_z} + \theta_{yz(ij)}(\omega)) \\ \hline |t_{zx(ij)}(\mathrm{i}\omega)|F_{j_x}\cos(\omega t + \phi_{j_x} + \theta_{zx(ij)}(\omega)) + \\ |t_{zy(ij)}(\mathrm{i}\omega)|F_{j_y}\cos(\omega t + \phi_{j_y} + \theta_{zy(ij)}(\omega)) + \\ |t_{zz(ij)}(\mathrm{i}\omega)|F_{j_z}\cos(\omega t + \phi_{j_z} + \theta_{zz(ij)}(\omega)) \end{bmatrix} .\end{aligned} \tag{4.10}$$

In fact, $\boldsymbol{y}^{ss}_{(i)}(t)$ defines a sinusoidal motion in all coordinate axes with different amplitude and phase, but at the same frequency.

4.2.1 PAC versus static perturbation methods

When PAC analysis is run with a static, time-invariant excitation, i.e., one that is at zero frequency, one can obtain the formulation used in static perturbation analysis. PAC with zero frequency perturbations is called as PDC, following the terminology used in electrical circuit simulations (AC/DC). Here, we show that static perturbation methods are in fact equivalent to PDC. However, for this equivalence to hold, the following should be noted for the static methods:

- ANM and/or MD based simulations are used to compute and form the Hessian of the potential energy function, and the N particles usually correspond to the C_α atoms of the residues.

- The excitation force is not time dependent, it is static $\boldsymbol{F}_e(t) = \boldsymbol{F}_e$, hence the perturbation-response analysis is essentially conducted at zero frequency: $s = 0$.

- The masses of the residues are usually assumed to be identical: $\boldsymbol{M} = \boldsymbol{I}$ and a loss/friction term is not incorporated into the equations: $\boldsymbol{L} = \boldsymbol{0}$.

- The displacements of the residues are taken as output observables. For example, the displacement of the i^{th} residue can selected by setting $\boldsymbol{C}_1 = \boldsymbol{I}$, $\boldsymbol{C} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \end{bmatrix}$ and then selecting the corresponding entries in the $\boldsymbol{y}$ vector, $\boldsymbol{y}_{(i)} = \begin{bmatrix} \boldsymbol{y}_{3i-2}, & \boldsymbol{y}_{3i-1}, & \boldsymbol{y}_{3i} \end{bmatrix}^T$.

In PDC, and hence in static perturbation methods, the frequency of the excitations is simply set to zero. Then, the transfer function $\boldsymbol{T}$(s) boils down to the following

$$\boldsymbol{T}(s=0) = -\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B} \ . \tag{4.11}$$

With $\boldsymbol{L} = \boldsymbol{0}$ and $\boldsymbol{M} = \boldsymbol{I}$, and

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \boldsymbol{0} & -\boldsymbol{H}^{-1} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} \ , \tag{4.12}$$

the transfer function at zero frequency, $\boldsymbol{T}(s=0)$, becomes

$$\boldsymbol{T}(s=0) = \begin{bmatrix} \boldsymbol{C}_1 & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{H}^{-1} \\ \boldsymbol{0} \end{bmatrix} = \boldsymbol{C}_1\boldsymbol{H}^{-1} \ . \tag{4.13}$$

Then, the displacements at zero frequency can be computed by multiplying the transfer function with the external static force vector

$$\boldsymbol{Y}(s=0) = \boldsymbol{C}_1\boldsymbol{H}^{-1}\boldsymbol{F}_e \ . \tag{4.14}$$

One can observe any element of the displacement vector, $\Delta\boldsymbol{r}$, by choosing $\boldsymbol{C}_1 = \boldsymbol{I}$. The entire displacement vector is simply given by

$$\Delta\boldsymbol{r}(s=0) = \boldsymbol{H}^{-1}\boldsymbol{F}_e \ . \tag{4.15}$$

We note that $\Delta\boldsymbol{r}(s=0)$ in Equation 4.15 is in fact equal to $\Delta\boldsymbol{r}$ in Equation 2.13, provided that the Hessian matrix, $\boldsymbol{H}$, is obtained via the same procedure. Hence, PAC can be considered to be a generalized form of the static perturbation methods,

extending it to handle dynamic perturbations with nonzero frequency. In addition, including a viscous friction term represented by $\boldsymbol{L}$ and choosing $\boldsymbol{M}$ according to the residue masses further improves the modeling and analysis fidelity, as compared to static methods.

4.3 Frequency dependent metrics for proteins

The standard output of PAC analysis is in the form of complex valued vibration amplitudes in 3D coordinates for each particle of the system under an input excitation. In their raw forms, the results are difficult to visualize and interpret for a protein, since the response of each system particle is represented in a six dimensional space, resulting from one complex valued amplitude for every coordinate axis. The analysis results need to be summarized in such a way so that the sought relationship between the input and the output becomes apparent. When one tries to quantify and summarize the response to a perturbation, there is a delicate balance between reducing complexity and preserving essential information. Protein analysis and characterization requires useful and concise metrics to be developed that can quantify local response behavior [59]. Our aim here is to provide scalar valued metrics for each of the residues or system particles that can be easily visualized with 2D plots. We next define several frequency dependent metrics in order to characterize the response of a particle.

Let $\Delta\boldsymbol{r}_{(i)}(\omega) = [\Delta x_i,\ \Delta y_i,\ \Delta z_i]^T$ denote the complex valued displacement vector for the i^{th} particle, that is computed by PAC.

Displacement magnitude, $|\Delta\boldsymbol{r}_{(i)}(\omega)|$, is the simplest and most widely used metric that can be computed with

$$|\Delta\boldsymbol{r}_{(i)}|(\omega) = \sqrt{\Delta {x_i}^2 + \Delta {y_i}^2 + \Delta {z_i}^2}\ . \tag{4.16}$$

At zero frequency (PDC), this metric can be used to determine the "most effective" input excitation direction and in order to compute correlations with experimental data. Moreover, one can use this metric to analyze how frequency of excitation affects the vibratory displacement magnitudes of the residues.

Energy, $E_i(\omega)$, given by

$$E_i = \frac{1}{2} m_i |\Delta \dot{\boldsymbol{r}}_{(i)}|^2, \text{ where } |\Delta \dot{\boldsymbol{r}}_{(i)}| = \omega |\Delta \boldsymbol{r}_{(i)}| \tag{4.17}$$

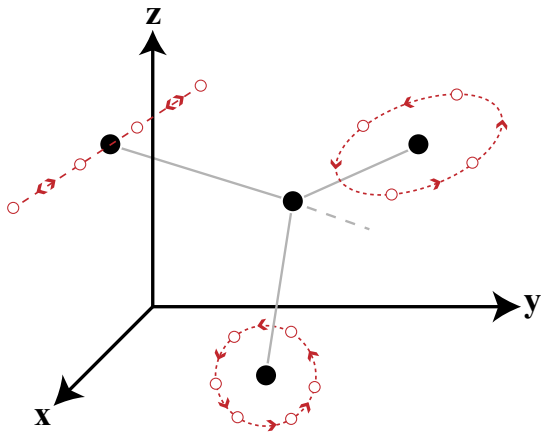
is the kinetic energy calculated using the velocities of the particles. Since frequency domain analysis is now possible with PAC and energy is a highly frequency dependent quantity, PAC enables internal energy transfer analysis for a protein. How energy of each particle changes with the excitation frequency provides insight on the properties and functions of the particles. Highly energetic particles can be considered as *critical* for the protein in ligand binding. Additionally, energy distribution in a protein under various types of environmental noise and disturbances can be analyzed.

At steady-state, the displacement vector $\Delta \boldsymbol{r}_{(i)}(\omega)$ that is obtained using Equation 4.10, by choosing the output as the displacements of all of the particles, i.e., $\boldsymbol{C}_1 = \boldsymbol{I}$, defines a parametric equation for an elliptical trajectory. Generally, the phase constants of the excitation, $\phi_{i_{x,y,z}}$, are set to zero for simplicity. In order to compute the actual absolute coordinates of the particles, the expansion point, i.e., the native state coordinates, are added to the deviation values

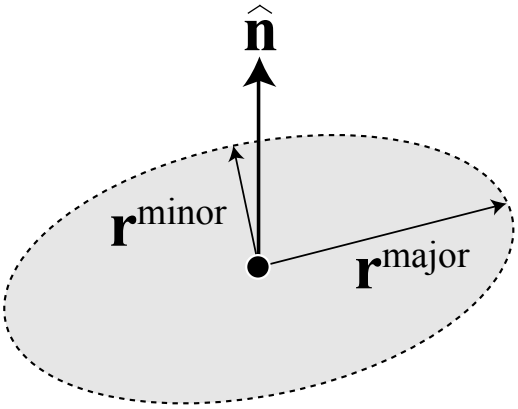
$$\boldsymbol{r}_i(t) = \Delta \boldsymbol{r}_{(i)} + \bar{\boldsymbol{r}}_i \ . \tag{4.18}$$

Semi-major and semi-minor axes of the elliptical trajectory of the i^{th} particle is defined by the vectors $\boldsymbol{r}^{major}{}_i(\omega)$ and $\boldsymbol{r}^{minor}{}_i(\omega)$, respectively. The lengths of the semi-major and the semi-minor axes are $|\boldsymbol{r}^{major}{}_i(\omega)|$ and $|\boldsymbol{r}^{minor}{}_i(\omega)|$, while the unit directions are $\hat{\boldsymbol{r}}_i^{major}(\omega)$ and $\hat{\boldsymbol{r}}_i^{minor}(\omega)$. Figure 4.1(a) shows time trajectories for three virtual particles, where one particle follows a linear trajectory, the other one follows a circular path, and the last one moves on a typical elliptical trajectory. The characteristics of a particle's trajectory varies with the frequency of excitation, which may help interpret and decipher the dynamics within the protein.

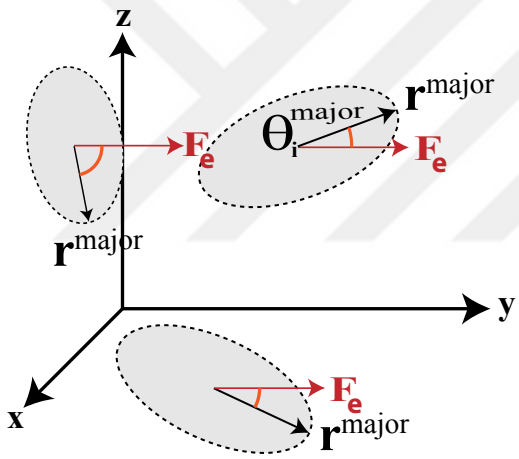
Major-axis length, $|\boldsymbol{r}^{major}{}_i(\omega)|$, is a measure of how far the particle directionally moves from the native state with the force excitation. This is different from the displacement magnitude metric, $|\Delta \boldsymbol{r}_{(i)}(\omega)|$. The maximum distance between two points of the elliptical trajectory is captured by the major-axis length metric.



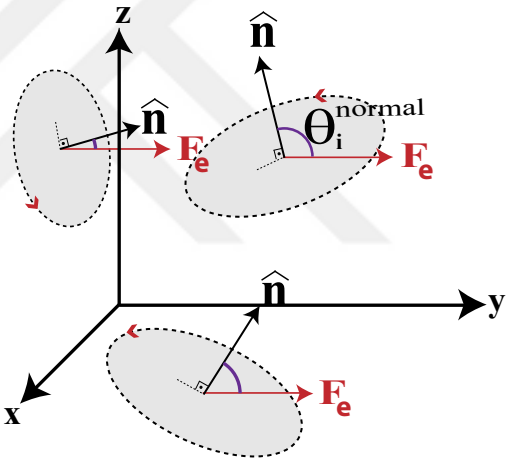
(a) Time trajectories for three different particles: linear, circular, and elliptical



(b) Semi-major axis, semi-minor axis, and the plane normal vector



(c) Angle between the major-axis and the force: $\theta^{major}{}_i$.



(d) Angle between the plane normal and the force: $\theta^{normal}{}_i$.

Figure 4.1: Visualization of the metrics

Angle between the major-axis and force, $\theta^{major}{}_i(\omega)$,

$$\theta^{major}{}_i = \cos^{-1}(|\hat{\boldsymbol{r}}_i^{major} \cdot \hat{\boldsymbol{F}}_e|) \tag{4.19}$$

is the angle between the applied force direction, $\hat{\boldsymbol{F}}_e$, and the semi-major-axis direction, $\hat{\boldsymbol{r}}_i^{major}(\omega)$. Among the two semi-major axes directions ($\hat{\boldsymbol{r}}_i^{major}(\omega)$ and its opposite $-\hat{\boldsymbol{r}}_i^{major}(\omega)$) the one that has a smaller angle with the force is selected,

i.e., ${\theta^{major}}_i(\omega)$ takes values between 0 and $\pi/2$, where smaller ${\theta^{major}}_i(\omega)$ values correspond to aligned motion with the force and larger values represent orthogonal motion. The angle between the major-axis and the excitation force direction may reveal hinge residues and co-moving groups of residues.

Isotropicity, $\xi_i(\omega)$, defined by

$$\xi_i = \frac{|{\boldsymbol{r}^{minor}}_i(\omega)|}{|{\boldsymbol{r}^{major}}_i(\omega)|} \tag{4.20}$$

is the ratio of the minor-axis and the major-axis lengths. It is used to distinguish the anisotropic particles that have a dominant direction of oscillation from the isotropic ones whose oscillatory motions reside in a plane without a significant dominant direction. For a particle with large displacement magnitude and/or energy, isotropicity may indicate different roles for that particle in the protein. For example, isotropic particles may serve as hubs to distribute energy to paths with different directions, whereas anisotropic ones may serve critical roles on a certain path. The major-axis directions of the isotropic particles are not as informative as the major-axis directions of the anisotropic ones. As such, co-consideration of isotropicity and the major axis direction is useful. Isotropicity, ξ, takes values between 0 and 1 for all possible trajectories: 1 for circular trajectories and 0 for linear ones. The particles with smaller values are more anisotropic than the others.

Angle between the plane normal and force, ${\theta^{normal}}_i(\omega)$, defined by

$${\theta^{normal}}_i = \cos^{-1}(\hat{\boldsymbol{n}}_i \cdot \hat{\boldsymbol{F}}_e) \tag{4.21}$$

is the angle between the applied force direction, $\hat{\boldsymbol{F}}_e$, and the normal of the plane containing the elliptical trajectory, $\hat{\boldsymbol{n}}_i = \hat{\boldsymbol{r}}_i^{major} \times \hat{\boldsymbol{r}}_i^{minor}$. ${\theta^{normal}}_i(\omega)$ is $\pi/2$ for the i^{th} particle when its elliptical trajectory is co-planar with the applied force. As θ^{normal} deviates from $\pi/2$ to 0 or π, the force direction deviates more and more from the trajectory plane. Please note that the major- and minor-axes directions appear in a certain order in the definition so that the ambiguity in the normal direction of the plane is removed.

4.4 PAC analysis of ligand (un)binding

For the rest of our treatment, we will adopt the C_α based coarse-grained modeling approach, and hence the system particle corresponds to the C_α atom of each residue. Although, with PAC, one can compute the displacements and the velocities of all of the residues under various input excitations with varying directions, magnitudes, and the frequency of the external force, not all input configurations may have a correspondence in real protein-ligand interactions. In this section, we review the implementation details that are mostly followed from [22], regarding the selection of appropriate input configurations for PAC analysis of ligand binding-unbinding for proteins.

4.4.1 Input residue(s) selection

An input residue is defined as the residue upon which an external force is applied. When investigating (un)binding, the input residue candidates are usually the binding site residues. If the binding site residues are known for a protein, they can be directly used. Any one (a group or all) of the residues from the binding site can be selected as the input residue(s). By perturbing all of the residues of the binding site, specific input residue(s) can be identified as the one(s) that exhibit the most correlated transient, dynamic behavior with the ligand (un)binding behavior. If both ligand-free and ligand-bound coordinates are available in the RCSB Protein Data Bank [60], PDC analysis can be used to determine the most appropriate input residue. The residues of the binding site are excited individually with a constant external force with the best direction and magnitude (please see Section 4.4.2). The one(s) that exhibit the highest correlation between the computed displacements and the experimental ones (between the ligand-free and ligand-bound forms) can be selected as the input residue.

4.4.2 Force magnitude and direction selection

PAC analysis methodology depends on whether or not the protein has both ligand-free and ligand-bound PDB forms. The procedure for force magnitude and direction selection is summarized in Figure 4.2.

For the proteins for which both ligand-free and ligand-bound forms are known, the procedure in Figure 4.2 is used to analyze binding (for unbinding, ligand-free and ligand-bound forms should be swapped). First, the Hessian is constructed for the ligand-free form based on ANM for the coarse-grained model, or by Taylor's expansion of the force field around the native state. Then, the ligand-bound form is aligned to the ligand-free form using the superpose function in MATLAB [61]. Other alignment tools can also be employed for this purpose. Any constraint can be imposed during alignment if there is prior knowledge on the fixed parts of the protein. Experimental displacements are computed as the difference of the C_α positions of the ligand-free and the aligned ligand-bound forms. Unit sphere (radius $r = 1$) is uniformly sampled in azimuth, ϕ, and zenith, θ, angles of the spherical coordinate system. For each of the sampled directions, a constant (time-invariant, at zero frequency) external force is applied to a selected input residue, and the displacements computed via PAC are post-processed as follows. Residue displacements are added to the coordinates of the ligand-free form and a new coordinate file is obtained. Then, this coordinate file is also aligned to ligand-free form and a computed ligand-bound form is obtained. By subtracting the C_α coordinates of ligand-free form and the computed ligand-bound form, computed displacement vector is obtained for every residue. Then, Pearson correlation coefficient[2] between this computed displacement vector, $\boldsymbol{r}^c$ and the experimentally determined one, $\boldsymbol{r}^e$, is calculated for

[2]We would like to note that, throughout the thesis, the correlation coefficient is used in order to figure out whether there is a causality relationship between two quantities. The correlation is computed by simply averaging over all of the residues in order to arrive at an overall characterization for the protein. It is not an actual statistical correlation computed over an ensemble of statistical samples.

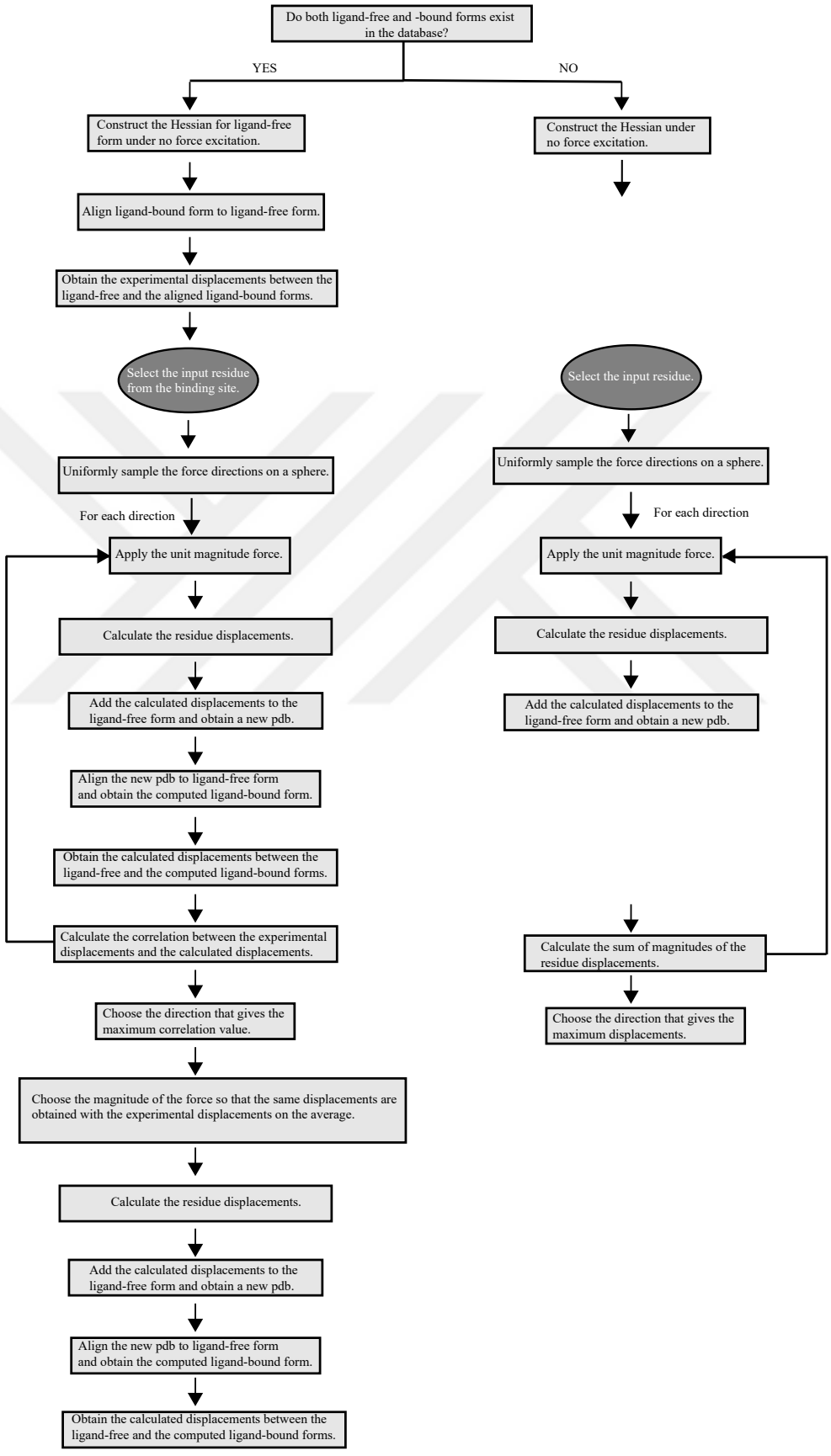


Figure 4.2: Force magnitude and direction selection procedure for proteins with both the ligand-free and the ligand-bound forms, and with only one form available.

each perturbed residue i as follows

$$corr_i = \frac{\sum_{j=1}^{N} (\boldsymbol{r}_j{}^e - \overline{\boldsymbol{r}^e})(\boldsymbol{r}_j{}^c - \overline{\boldsymbol{r}^c})}{\sqrt{\sum_{j=1}^{N} (\boldsymbol{r}_j{}^e - \overline{\boldsymbol{r}^e})^2} \sqrt{\sum_{j=1}^{N} (\boldsymbol{r}_j{}^c - \overline{\boldsymbol{r}^c})^2}} , \tag{4.22}$$

where overbar indicates the average over the residues. For each sampled direction, this procedure is repeated. Then, the direction that gives the highest correlation is selected as the force direction. Force magnitude is chosen so that the same displacement values are obtained on the average. After determining the force magnitude and the direction, residue displacements are recomputed with the determined force parameters. The *computed ligand-bound form* is finally obtained after an additional alignment step to the ligand-free form. In frequency sweep analysis via PAC, the force direction determined as such at zero frequency is kept. The static force magnitude determined at zero frequency is used as the amplitude of the time-varying force excitations at nonzero frequencies. PAC analysis provides insight into the role of oscillatory, time-varying excitations in the (un)binding process.

If both apo and holo forms are available, but the binding site is not known, then one can test all of the residues with a force excitation. For each residue, the force direction and magnitude can be determined following the protocol described in Figure 4.2. Force direction search on a spherical grid for each residue will incur some computational cost. But this will still be fast when compared with time-domain MD simulations. The major advantage of the proposed PAC formalism stems from its computational efficiency. Hence, it can be used for high throughput screening purposes, in cases when the binding residues are not known.

If both PDB forms of the protein are not available, the alignment and magnitude scaling steps are skipped and the force direction that results in the maximum sum of magnitudes of the computed residue displacements is determined. This perturbation direction determination procedure can be applied whether the input residue is already known or is to be determined. If there is no information on the binding site and/or new possible binding sites are sought, the most effective perturbation direction is determined as above for all candidate residues. Then, the resulting dis-

placements with the determined effective perturbation direction for each residue are examined in order to identify the key residues that may belong to a binding site. In comparing perturbation directions and input residues, the force magnitude is set to unity. Since PAC analysis is based on a linearized model, the ratio of the output response to the input magnitude is independent of the force magnitude used. Here, we have outlined a plausible methodology for the determination of effective perturbation directions and/or candidate input residues. This methodology can be improved and extended. For instance, input perturbations that excite multiple residues simultaneously may serve as a better model for ligand binding. The procedure for proteins with only one form available is also summarized in Figure 4.2.

4.5 Results

4.5.1 Setup and preliminaries

We first provide information on the setup that was used to produce the results that are reported:

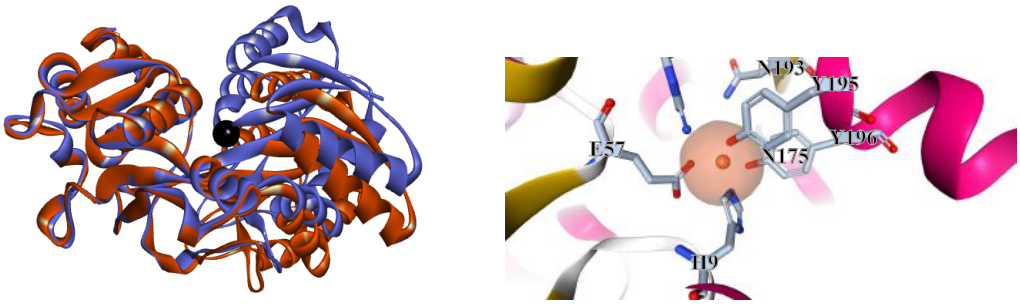
3D coordinates of the proteins are retrieved from the Protein Data Bank [60, 62]. The Hessian is constructed based on an ANM model with a pairwise interaction cut-off distance value, r_c, of 8 Å for the ferric binding protein (FBP), 10 Å for the bacterial chemotaxis protein (CheY). In determining these values, the cut-off distance is swept within the 7–15 Å range and the value that produces the best correspondence between the computed and experimental displacements is chosen [63, 64, 53, 65]. Hookean spring constants are taken to be identical and set to 1 for all of the interacting residue pairs within the cut-off distance. Masses of the residues are taken into account throughout the analysis, and the mass matrix, $\boldsymbol{M}$, is constructed using the average mass values that are taken from [66]. The loss matrix, $\boldsymbol{L}$, is set to a scaled identity matrix $\boldsymbol{L} = 0.3\boldsymbol{I}$ so that the solvent effects are included in the model. With $\boldsymbol{L}$ chosen as such, the high frequency movements do not completely vanish, and at low frequencies, they are slightly subdued, when compared to the experimental zero frequency displacement magnitudes. The frequency range for PAC

analysis is set as the interval from zero to the half of the maximum normal mode frequency calculated for each protein. This frequency range is uniformly sampled at 200 points, where at every point a PAC analysis is run. In addition to these frequency points, normal mode frequencies are also included in the frequency set. The transfer function becomes singular at normal mode frequencies if loss/friction term is set to zero. Therefore, loss terms are always included for running PAC at the normal mode frequencies.

Two proteins (FBP and CheY) were analyzed with PAC following the protocol described in Section 4.4. FBP and CheY were chosen because they were considered in the original works in the PRS method [22]. Furthermore, for FBP, MD studies in the literature have shown that single-point static force application at a single residue models the ligand binding event quite well [22]. For both of these proteins, the ligand molecules (ions) are small. Therefore, a single-point force excitation model suffices to capture the ligand binding. As PAC extends the conventional static methods into the frequency domain, by the chosen proteins, we are able to compare the results obtained by PAC to the ones obtained in the literature with PRS. We show that we obtain the same results (up to some extent, possibly due to variations in the protocols used) with PDC as in PRS. We further these results that were reported in the literature by extending them with a dynamic, time-varying, frequency domain analysis that is enabled by PAC. We compute the frequency dependent metrics described in Section 4.3 via PAC, providing higher fidelity information on the perturbation response dynamics. The following section presents the results for the FBP. Please see Section 8.1.1 for our investigation of CheY protein.

4.5.2 Ferric binding protein (FBP)

FBP is a highly conserved protein in the iron acquisition system of gram-negative bacteria. It resides in the periplasm and receives the iron obtained from the mammalian host and transfers it to the pathogen [67]. Iron is an essential element that is acquired from the host and utilized in many essential biological processes including bacterial cell division [68]. Infection initiation requires the growth and colonization



(a) Apo form (1D9V) and holo form (1MRP) of FBP are aligned, apo form is colored with blue and holo form is colored with orange, both in solid ribbon form. Ligand (Fe^{3+} ion) is shown as a black circle. The plot was created with [74].

(b) Binding pocket of Fe^{3+} ion and inter-acting residues: H9, E57, N175, N193, Y195, and Y196. Plot was created with [75, 76].

Figure 4.3: Alignment of apo and holo forms and binding site of FBP

of a pathogen. Thus, FBP is a potential target for broad-spectrum antibacterial drugs against human pathogens such as *Haemophilus influenzae*, *Neisseria gonorrhoeae*, and *Neisseria meningitidis*[69]. Ligand-free (apo form with PDB ID: 1D9V) and ligand-bound (holo form with PDB ID: 1MRP) structures of FBP are available with a resolution of 1.75 Å and 1.6 Å respectively, in PDB. The aligned apo and holo forms are shown in Figure 4.3(a). FBP is a monomeric protein with 309 amino acids. FBP exhibits a 20° closure between two interacting domains about a central beta-sheet hinge upon binding of a ferric ion [70]. Presence of iron stabilizes the protein and closes the inter-domain hinge. The closest residues to Fe^{3+} are H9, E57, N175, N193, Y195, and Y196 and the ligand binding site can be seen in Figure 4.3(b). Interested reader may refer to [22], [71], [72], [73] for further information on the structure and function of FBP, and its iron transport mechanisms.

Displacement magnitudes, $|\Delta \boldsymbol{r}_{(i)}(\omega)|$, of C_α atoms are calculated after the alignment of the ligand-bound form to the ligand-free form. Alignment is done by

using only the fixed domain residues 102-225. RMSD between the two structures is calculated as 0.302 Å after alignment.

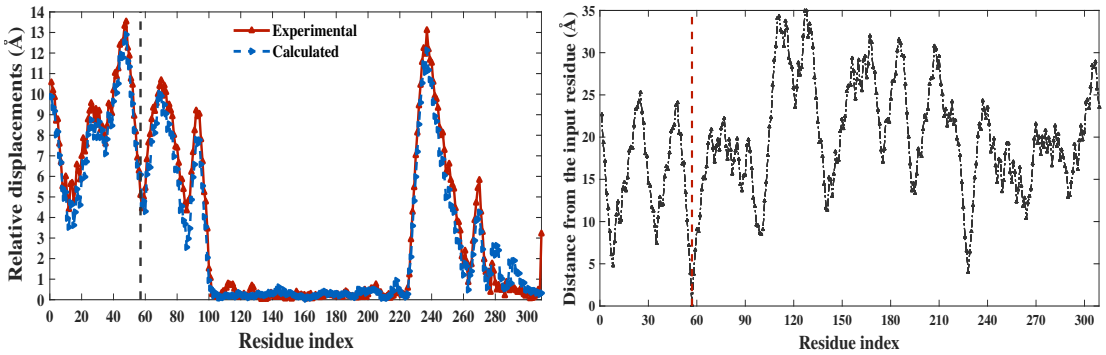
It was shown in [22] that the constant, time-invariant force perturbations upon only residue 57 (which mimics ion binding) results in highly correlated responses with MD simulations. Therefore, we have also selected residue 57 (Glu57) as the input excitation residue for PAC analysis.

First, experimental displacements for all of the residues are calculated simply by computing the differences of the coordinates for the experimentally determined apo and holo forms after alignment. Computationally, the holo (apo) form is obtained from the other one by calculating the displacements resulting from the constant input force excitation. In this case, one form (apo) is considered to be the initial unperturbed state and the other one (holo) results from the force perturbation. Next, the correlation between the experimental and computed displacements is used as an indicator of how well the input force exertion on the input residue can capture the ligand (un)binding event. Pearson correlations are calculated using Equation 4.22. The correlation coefficient between the experimental displacements and the calculated ones (after alignment) for FBP is found to be 0.9731. Table 8.1 summarizes the force excitation related parameters and the correlation values between the experimental and the computed displacements for FBP as well as CheY.

In Figure 4.4(a), the red curve shows the experimental displacements and the blue one is for the computed (calculated). The force direction and magnitude were chosen to produce the best correlation value, as described in Section 4.4.2. Figure 4.4(b) shows the distances of all of the residues to the input residue, residue 57.

The correlation value for FBP reported above is in fact better than the ones reported for LRT and PRS [26, 22]. Minor differences in the results may originate from the alignment tools, the selection of the fixed and moving domain residues for alignment, as well as the force magnitude and direction selection protocol. All of the methods, PRS, LRT and our PDC achieve correlation values around 0.95 with the respective best force parameters.

We now present PAC analysis results for dynamic force excitations, with the



(a) Experimental and calculated displacements between apo and holo forms of FBP. Constant force excitation on residue 57 successfully predicts ligand binding.

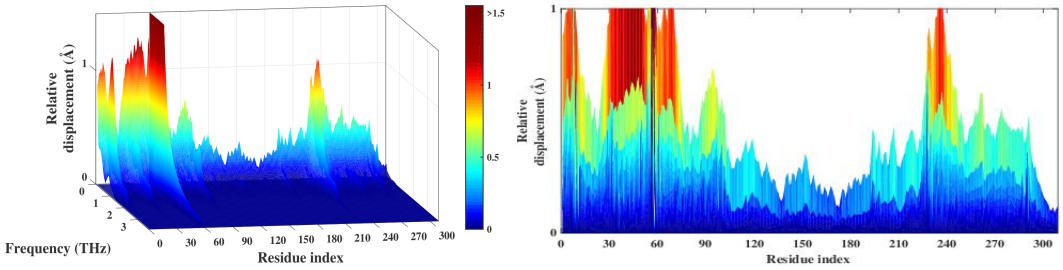
(b) Residue distances to the input residue (residue 57)

Figure 4.4: Displacement magnitudes of FBP residues in Å. Red curve shows experimental displacements and blue curve is for the computed displacements at the best force magnitude and direction at zero frequency. Residues 8, 9, 55, 56, 58, 59, 227, 228, and 229 are within 8 Å distance of residue 57.

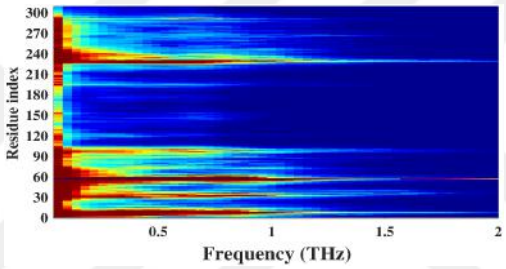
input residue, force magnitude and direction set to the same ones used for the static excitation case described above. Figure 4.5(a) shows the displacement magnitudes as a function of both the residue index and the frequency of excitation. The same data (3D plot) is presented in Figures 4.5(b) and (c) from two other view angles. For better visualization purposes, the displacement values above 1.5 Å for the input residue are clipped. Displacement magnitudes as functions of both the residue index and frequency reveal complex and rich characteristics. At most of the frequency intervals, almost all of the residues follow a similar trend, but for particular frequency intervals, some residues behave much differently than the others. The residues that exhibit larger displacements are 8-10, 55-60, and 226-235. Higher frequency excitations (around frequency values 0.5 - 0.95 THz) result in an increase in the displacements for these residues. However after some residue-specific frequency threshold, further increase in the excitation frequency results in smaller magnitude

movements. The frequency at which each residue exhibits the maximum displacement magnitude is a unique property of that residue. Most of the residues exhibit their maximum displacements at low frequencies. However, residues 8, 55, 56, 57, 58, and 227 reach their maximum value in the mid frequency range. These are among the closest residues to the input residue. However, not all residues that are close to the input residue display this behavior. Figure 4.5(d) shows the displacements for a set of selected residues, 8, 9, 44, 48, 56, 58, 193, 196, 230, and 290. Residues with pronounced bimodal characteristics are 9 and 56. They are relatively closer to the input residue. However, not all of the residues that are close to the excitation site display a bimodal characteristics. Bimodal characteristics could be an important indicator of frequency selectivity and a sort of multi-resonant behavior. The frequencies that correspond to the peak points of the characteristics can be regarded as the most effective frequencies through which the protein can interact with other proteins or ligands. One can also surmise that interactions through residues that exhibit bimodal characteristics have a more complex nature, pointing to a multi-frequency nature of the interaction. Residue 230 exhibits less displacement than residue 9 at low frequencies, but in the higher frequency range, the situation is reversed. Some of the residues (e.g., 8, 9, 56, 58, 290) show a drastic increase in the mid-frequency range but others do not (e.g., 193, 196). Figure 4.5(e) shows the displacements for all of the residues at five selected frequencies in logarithmic scale.

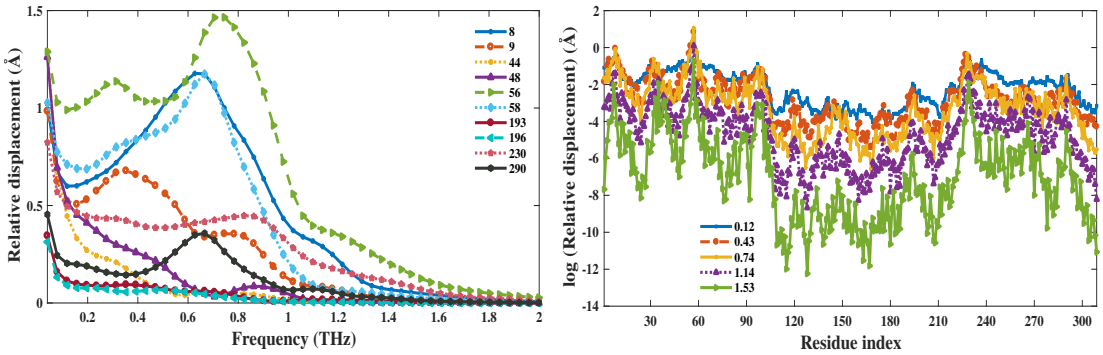
Kinetic energy, $E_i(\omega)$, calculated at each frequency, is proportional to the square of both the magnitudes of frequency-dependent displacements and the frequency, essentially their product. For a given frequency, larger amplitude motions have higher kinetic energy. For a given displacement amplitude, higher frequency motions have higher kinetic energy. The excitations at low frequencies result in relatively larger amplitude fluctuations, due to the low pass nature of the frequency response of a protein. However, the high value of displacement magnitude is counteracted by the low frequency value in the product. On the other hand, higher frequency excitations result in lower amplitude motions, counteracting the high fre-



(a) Residue displacements in 3D view (b) Residue displacements vs residue index



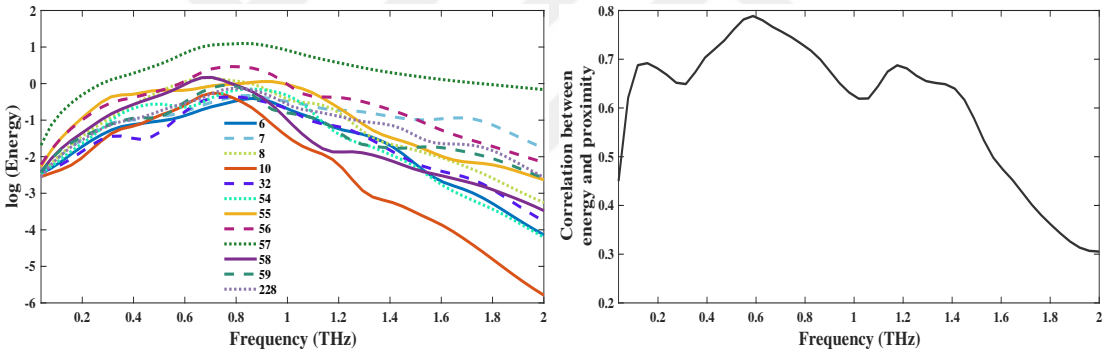
(c) Residue index vs frequency



(d) Relative displacements of selected residues (e) Relative displacements of all residues for selected frequencies

Figure 4.5: Relative displacement magnitudes as a function of residue index and frequency

quency value in the product. That is, at both low and high frequencies, the kinetic energy has a relatively small value due to the opposing terms (displacement and frequency) in the product. As expected, Figure 4.6(a) shows that the residues have higher kinetic energy for mid-range frequency excitations. It is of course expected that the input residue has the maximum energy. However, some of the non-input residues exhibit more energy variation than the others. The residues that show drastic energy variations can be considered as the *critical* ones for energy transmission within the protein. It is worth pointing out that the critical residues attain their maximum energy values at distinct frequencies. Therefore, the frequency of the force excitation is a determining factor on the energy distribution of the residues. Figure 4.6(a) shows the logarithmic scale energy plots for these *critical* residues. A



(a) Energy of *critical* residues

(b) Correlation between energy and proximity to the input residue as a function of frequency

Figure 4.6: Energy analysis of FBP

residue is regarded as *critical* if the maximum energy value of that residue is greater than or equal to 3% of the maximum energy value that any residue attains over the whole frequency range in the protein. This is typically the maximum energy of the input residue, however, there may be cases where a residue other than the input residue may attain the maximum energy. With this criterion, the critical residues of FBP are 6, 7, 8, 10, 32, 54, 55, 56, 57, 58, 59 and 228. It is expected that some

of these critical residues are the closest ones to the input residue, since that is the entry point for energy into system. However, what is worth pointing out is that some of the residues nearest to the input are not among the critical, and some of the distant residues are critical instead. The distant critical residues may be involved in allosteric communication.

Let $\boldsymbol{r}_{(j)}$ denote the coordinates of the j^{th} residue in the initial form, i.e., the holo form after alignment. Then, with i^{th} residue as the input, the proximity of the other residues to the input site are calculated as the inverse of the distance between the i^{th} and the j^{th} residue.

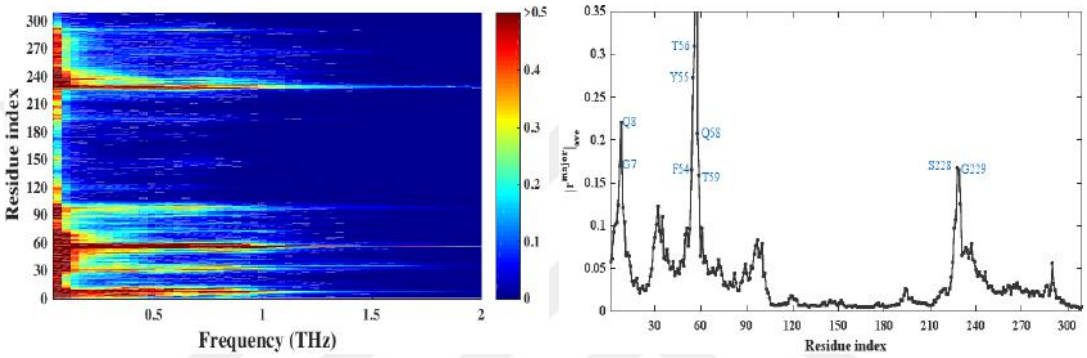
$$\epsilon_{ij} = \frac{1}{|\boldsymbol{r}_{(i)} - \boldsymbol{r}_{(j)}|} \;\; j = 1 \ldots N. \;\; j \neq i. \tag{4.23}$$

We can examine how energy relates to how proximate the residue is to the force application site with a Pearson correlation calculated as follows

$$\begin{aligned} corr(\omega) &= \frac{\sum_{j\neq i}^{N}(\epsilon_{ij} - \overline{\epsilon_i})(E_j(\omega) - \overline{E(\omega)})}{\sqrt{\sum_{j\neq i}^{N}(\epsilon_{ij} - \overline{\epsilon_i})^2}\sqrt{\sum_{j\neq i}^{N}(E_j(\omega) - \overline{E(\omega)})^2}} \text{ with} \\ \overline{\epsilon_i} &= \sum_{j\neq i}^{N} \epsilon_{ij} \Big/ (N-1) \;, \end{aligned} \tag{4.24}$$

where $\overline{\epsilon_i}$ is the average of the proximity and $\overline{E(\omega)}$ is the average energy over the residues. Figure 4.6(b) shows the correlation between proximity to the input residue (see Figure 4.4(b)) and energy as a function of the excitation frequency. In certain frequency intervals (0.15-0.25, 0.3-1, 1.1-1.6 THz), the energy has a higher correlation with proximity, whereas for low and high frequencies they are relatively uncorrelated. It is also interesting that there exist distinct frequency values where the correlation exhibits peaks. That is, the correlation of proximity to the input site and energy distribution does not follow a monotonic trend with frequency. The uncorrelated frequency intervals can be considered as an indication of allosteric behavior. Allosteric behavior may not be a statically determined property of a protein. Instead, the dynamic, time-varying characteristics of the excitation may play an important role in allostery.

Major-axis length, $|\boldsymbol{r}^{major}{}_i(\omega)|$, reveals the extent of the range of the motion of the residues, as a function of excitation frequency. In Figure 4.7, the major-axis length is shown both as a function of residue index and frequency, and by averaging over frequency. Based on the frequency averaged results, residues 7, 8, 54-59, 228, and 229 move more than the others. Those that exhibit the least amount of movement are the region between 130-190, and residues 307-308.



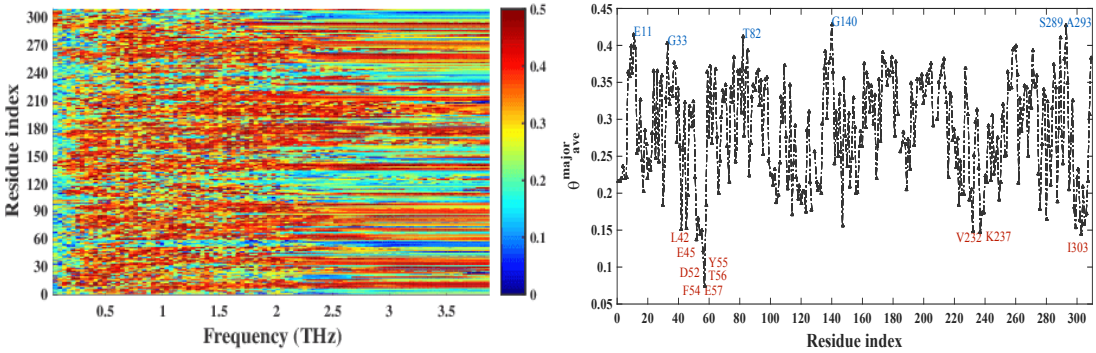
(a) $|\boldsymbol{r}^{major}|$ (b) Frequency-averaged $|\boldsymbol{r}^{major}|_{ave}$. Values above 0.35 Å are not shown.

Figure 4.7: Major-axis length $|\boldsymbol{r}^{major}|$ of FBP.

Angle between the major-axis and applied force direction, $\theta^{major}{}_i(\omega)$, classifies the residues that move together in the same direction with the force, in the opposite direction, or in an orthogonal manner. $\theta^{major}{}_i(\omega)$ can help detect the hinge points in the protein.

Figure 4.8 shows that most of the residues tend to move in an orthogonal manner to the force for larger frequencies. Frequency averaged θ^{major} values are shown in Figure 4.8(b). On average, residues 11, 33, 82, 140, 289, and 293 move orthogonally to the force. Residues 42, 45, 52, 54, 55, 56, 57, 232, 237, and 303 are aligned with the excitation force direction.

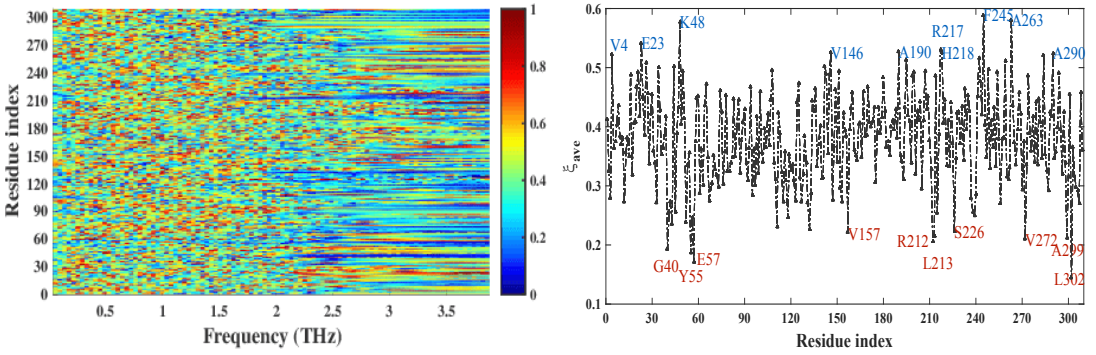
Isotropicity, $\xi_i(\omega)$. Residues with $\xi_i(\omega)$ values around 0 are *anisotropic*, whereas the ones with values near 1 are *isotropic*. Motion of the anisotropic residues indicates directional selectivity, such as opening/closing. Therefore, the anisotropicity



(a) θ^{major} (b) Frequency-averaged ${\theta^{major}}_{ave}$

Figure 4.8: Angle between the major-axis and force, θ^{major}, of FBP. Colorbar labels are to be multiplied with π, and in radians.

metric may be useful in the directional analysis of protein motion. In general, the isotropicity of a residue varies greatly as a function of frequency, although some of the residues may exhibit similar trends within all (some) of the frequency range.



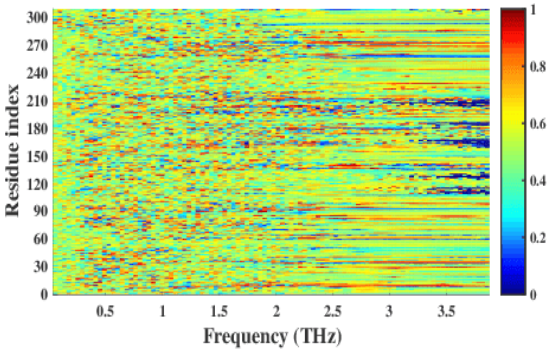
(a) ξ (b) Frequency-averaged ξ_{ave}

Figure 4.9: Isotropicity ξ of FBP.

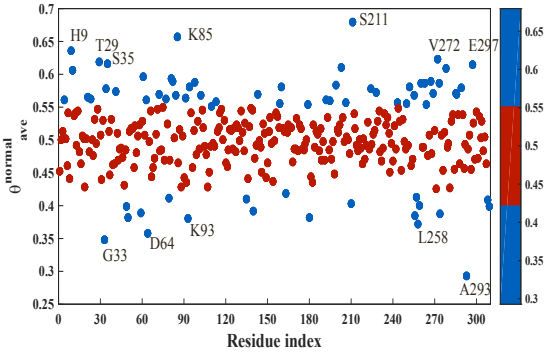
Figure 4.9(a) shows isotropicity as a function of frequency. It is highly frequency dependent for FBP. Over a certain frequency range, most of the residues become anisotropic. Therefore, we can claim that excitations with higher frequency force the protein motion to be more directionally selective. Residues 57, 67, 68, 73, 95, 111,

113, 115, 127, 163, 166, 167, 186, 210, and 302 have isotropicity values less than 0.03, at a certain frequency, while for residues 21, 23, 98, 124, 172, 179, 200, 202, 218, 219, 245, 254, 263, 281, and 284 it becomes more than 0.8. Figure 4.9(b) shows the frequency-averaged values of isotropicity and provides an overall impression about the residue on being (an)isotropic. Residues 40, 55, 57, 212, 272, and 302 are the most anisotropic ones on the average, and 4, 23, 48, 146, 190, 217, 218, 245, 263, and 290 are the most isotropic.

Angle between the plane normal and force, $\theta^{normal}(\omega)$. Figure 4.10 shows that, as the frequency increases, the plane in which most of the residue trajectories reside tend to be aligned with the force direction. There are exceptions to this, some residues move orthogonally to the force. These residues may be involved in directional motions which are functionally crucial within the protein. Frequency-averaged θ^{normal} plot in Figure 4.10(b) shows also that most of the residues are aligned with force, e.g., 5, 17, 18, 20, 57, 62, 83, 89, 207, 303, and 306. However, the motion of the residues 9, 29, 35, 33, 64, 85, 93, 211, 258, 272, 293, and 297 deviate from the force direction.



(a) θ^{normal}



(b) Frequency-averaged ${\theta^{normal}}_{ave}$. Red circles indicate the residues whose motions are aligned with the force, and blue circles show the orthogonal motions.

Figure 4.10: Angle between the plane normal and the force direction, θ^{normal}, of FBP. Colorbar labels in radians/pi.

Chapter 5

ALLOSTERY IN PROTEINS AS POINT-TO-POINT TELECOMMUNICATION IN A NETWORK: FREQUENCY DECOMPOSED SIGNAL-TO-NOISE RATIO AND CHANNEL CAPACITY ANALYSIS

5.1 Introduction

The coupled dynamics of biological molecules underlie essential functions in cellular processes, including protein regulation and cellular signaling. Allostery—the transmission of the effect of ligand binding to another (often) distal site of a protein—plays an important role in many of these functions and may also be leveraged for novel drug delivery applications [80, 81]. There has been sustained activity in both computational and experimental allostery research ever since Monod and Jacob introduced the concept [82]. Nevertheless, a general mechanistic understanding of allosteric processes still remains elusive [83].

Since its introduction, the definition of allostery has evolved over time. In the earlier structure-centric models, the ligand induced change in the binding affinity at a distal site was thought to be accompanied by significant conformational changes. Two historically dominant models for allosteric mechanisms are the Monod–Wyman–Changeux (MWC) [84] and the Koshland–Nemethy–Filmer (KNF) models [85]. Both models assume that there is an equilibrium between two pre-existing conformational configurations, corresponding to the active and inactive states, but differ in the treatment of intermediate states. In the MWC model, all subunits simultaneously transition from the inactive to the active configuration upon ligand binding,

The content of this chapter is under revision.

i.e., they undergo a *concerted* motion. On the contrary, in the KNF model, the subunits transition, one at a time, in a *sequential* manner, resulting in transition states. Despite their initial success, these structure-based models are now considered insufficient for explaining the behavior of some allosteric proteins, such as the PDZ domains. As a result, a variety of extended models have been proposed. For instance, in the population-shift model, instead of only two states, proteins are assumed to exist in an ensemble of conformations. Subsequent to ligand binding, the ensemble undergoes a population shift towards a state that is favored by the ligand [86]. A more recent view called *dynamic allostery*, that was introduced by Cooper and Dryden, led to a paradigm shift in the notion of allostery by showing that, even in the absence of an observable conformational change in the mean structure, some proteins can exhibit allosteric behavior [87]. This result suggests that all proteins may be considered as intrinsically allosteric [88]. Cooper and Dryden further demonstrated via statistical thermodynamics based analysis that cooperative interaction free energies could deviate on the order a few kJ/mol upon ligand binding as a result of the changes in the frequency and amplitude of thermal fluctuations, with only a subtle change in the mean structure [87]. This result suggests that a frequency domain analysis of the fluctuations may provide quantitative insight into the detailed mechanisms of action-at-a-distance phenomena in allostery.

Motivated by the work of Cooper and Dryden, the present study proposes a set of novel computational methods for investigating allosteric processes by examining the effect of perturbations, due to both ligand binding and noise, over a range of relevant frequencies. To facilitate the scanning of perturbation frequencies, we represent the protein as a mass-spring network, as a graph of nodes connected by edges. The nodes of the network correspond to the constituting atoms, or coarse-grained beads (defined as a collection of atoms), and the edges represent the interactions between them. Noise forces are applied to all nodes, to model the thermal fluctuations that originate from the protein-solvent interactions. Similarly, the ligand-protein interaction is represented by external forces, but they only act on the nodes that are in close contact with the binding site. The response at a (distal) node selected as

the output is then determined by the noise forces acting on all of the nodes that are further shaped by the network dynamics, as well as the external forces representing the ligand. When "signal transmission" from a specific input to multiple output sites is considered, signal attenuation, and transmission reliability that is also affected by noise, may vary a great deal over the output sites. Accordingly, the emergence of allosteric response due to ligand binding may be attributed to the *Signal-to-Noise ratio* (SNR) characteristics measured at the output site. We put forward that, in order for an allosteric response to occur at a certain output site, the associated SNR value must be relatively high.

Instead of treating ligand binding as a static structural event, we capture the dynamic nature of the ligand-protein interactions via modeling the external forces as dynamic oscillatory excitations, as in Chapter 4, and as in related work [89]. While the excitation frequency is swept over a relevant frequency range, the frequency dependence of SNR is fully characterized. Instead of considering only the low-frequency (global) modes of motion, we argue that a full spectrum analysis is essential, especially in cases where SNR does not follow a monotonic trend as a function of frequency. In particular, noticeably higher SNR within a particular frequency band may be the key in identifying dynamic allosteric response.

Akin to network representation of proteins, frequency decomposed SNR based analysis further enables the conceptualization of allostery as a *point-to-point channel* in a multi-channel, networked, noisy communication medium, pointing to the direct link between SNR and *channel capacity* that was established by Shannon. In the communications and information theory framework, the ligand serves as a transmitter (source of information), while the whole protein acts as the noisy medium of communication, i.e., the channel, and the region where the response is probed can be regarded as the receiver (of information). As in all communication systems subject to noise, there is an upper limit for the error-free (reliable, effective) information transmission rate from the transmitter to the receiver, defined as the channel capacity. Based on the Shannon-Hartley theorem, the channel capacity is determined by the integral of SNR over the relevant frequency band [90]. As an extension to

frequency decomposed SNR analysis, we propose to utilize channel capacity as an aggregate (over frequency) and interpretable measure of robustness of the communication between the ligand binding site and the distal regions of the protein in the presence of noise.

Previous efforts attempting to decipher allostery have mostly concentrated on the identification of *critical* residues, i.e., residues whose activities are most affected by the binding event. We propose to add another layer to this analysis by anatomizing each residue's frequency decomposed SNR profile individually. Furthermore, we propose two different methods for characterizing the channel capacity, namely, the *per residue scan* and the *binding pocket excitation* schemes. These methods differ in terms of the number of nodes upon which the external force is applied, as well as with respect to the excitation characteristics. Each analysis aims to detect distinct features. In the per residue scan, while the external force is applied to each (input) node, one at a time, the channel capacity values are calculated at all of the remaining nodes as candidate outputs. Through this analysis, the input-output residue pairs that have the potential to exhibit allosteric coupling can be identified without relying on any prior site-specific binding information. Hence, this analysis can be used to discover novel targets for ligand binding. In the binding pocket excitation scheme, the external force is applied specifically to the residues that are known to be located in the binding pocket of a specific ligand, while the channel capacities for the rest of the protein residues are calculated. The aim here is to identify the residues that are most likely to allosterically interact with an already known binding event.

The utility of the proposed methods is presented for a representative single-domain allosteric protein that is known to display dynamic allostery, namely the third PDZ domain of the PSD-95 protein. For an overview of previous approaches applied to PDZ3, the reader is referred to [81]. This study aims to contribute to this growing area of research through novel frequency-decomposed SNR and channel capacity based analysis techniques, which provide complementary insight with respect to previous methods. The proposed techniques have the potential to offer unique perspectives in probing the responses to external excitations while also taking the

noise characteristics into account.

The chapter is organized as follows: Section 5.2 describes the theoretical foundations, and Section 5.3 presents the proposed and employed methods. Section 5.4 shows the results for the investigation of allostery in PDZ3. Supplementary material is provided in Section 8.2.

5.2 Theory

We describe the Signal-to-Noise ratio (SNR) and the channel capacity formulations in a step-by-step manner in the following subsections. Section 5.2.1 presents the Langevin formulation for the network dynamics driven by thermal noise and external forces. The stochastic properties of the thermal fluctuations are also discussed. Computing transfer functions at low frequencies involves the inversion of an ill-conditioned (nearly singular) matrix, arising from the degrees of freedom related to rigid body rotations and translations. In Section 5.2.2 we address this issue by reformulating the inversion of a matrix as a constrained least-squares problem (CLS), where the constraints are obtained from Eckart's conditions [91, 92]. The solution of the CLS problem yields the transfer functions needed. Section 5.2.3 shows that mean square fluctuations are independent of the friction coefficients due to the adapted Langevin formulation. Section 5.2.4 introduces the *power spectral density* (PSD) that characterizes both deterministic and stochastic signals in a frequency decomposed manner. Section 5.2.5 presents a general scheme that utilizes the transfer functions in order to compute the PSDs of the displacements (outputs) from the PSDs of the applied forces and noise excitations (inputs). We describe in Section 5.2.6 how the mean square fluctuations of node (atom or coarse-grained bead) positions can be computed using transfer functions and PSDs when the network is excited only by noise. This is accomplished by simply integrating the PSDs over all frequencies. We show that the results obtained as such match the ones that are computed using a well known technique [93] that uses the pseudo-inverse of the Hessian of the potential energy function. The equivalence of the two techniques not only validates the constrained transfer function and the PSD approach, but also

leads to the characterization of the fluctuations in a frequency decomposed manner. In Section 5.2.7, we define frequency decomposed SNR as the ratio of the displacement PSDs due to the deterministic external forces (modeling ligand binding) to the ones due to noise alone. Finally, channel capacity is defined and computed based on frequency decomposed SNRs.

5.2.1 Dynamics: Langevin formulation

Langevin formulation described in Equation 2.8 for the linearized internal forces is adapted to describe the protein dynamics

$$\boldsymbol{M}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\boldsymbol{\Delta r} + \boldsymbol{L}\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\Delta r} + \boldsymbol{H}(\bar{\boldsymbol{r}})\boldsymbol{\Delta r} = \boldsymbol{\xi}(t) + \boldsymbol{F_e}(t)\ , \tag{5.1}$$

where $\boldsymbol{\Delta r} = \boldsymbol{r} - \bar{\boldsymbol{r}}$ is the displacement from the equilibrium position ($\bar{\boldsymbol{r}}$) and $\boldsymbol{H}$ is the Hessian. $\boldsymbol{M} = \mathrm{diag}(\boldsymbol{M}_{(1)}, \boldsymbol{M}_{(2)}, \ldots, \boldsymbol{M}_{(N)})$ is a block diagonal mass matrix. The effect of frictional forces exerted by the viscous fluid is incorporated into the dynamics via $\boldsymbol{L}$, which is taken to be proportional to the velocity of the particles. $\boldsymbol{\xi}(t)$ captures the background noise forces due to random collisions of the Brownian particles. $\boldsymbol{F_e} = [\boldsymbol{F_{e_1}(t)^T}, \boldsymbol{F_{e_2}(t)^T}, \ldots, \boldsymbol{F_{e_N}(t)^T}]^{\boldsymbol{T}}$ is the external force vector of size $3N \times 1$. Hydrodynamic shielding effects are ignored, and thus $\boldsymbol{L} = \mathrm{diag}(\boldsymbol{L}_{(1)}, \boldsymbol{L}_{(2)}, \ldots, \boldsymbol{L}_{(N)})$ is a diagonal matrix of the friction coefficients. We consider the special case of spherical particles, and the frictional force is assumed to act isotropically on the particles, i.e., $\boldsymbol{L}_{(i)} = \mathrm{diag}(\gamma_i, \gamma_i, \gamma_i)$. The friction coefficient γ_i for particle i with radius a_i is given by Stoke's law as

$$\gamma_i = 6\pi\eta a_i\ , \tag{5.2}$$

where η is the dynamic viscosity, and the hydrodynamic radius is calculated from the volume of the particle, V_i, as $a_i = (\frac{3V_i}{4\pi})^{1/3}$.

In the Langevin formulation, the noise and the viscous friction terms are linked to each other via the fluctuation-dissipation theorem [94, 95]. Particle positions continually fluctuate due to the interplay between the noise and the friction in the system, even in the absence of external force excitations, i.e., when $\boldsymbol{F_e} = \boldsymbol{0}$. $\boldsymbol{\xi}(t)$

in the Langevin equation is a wide sense stationary (WSS) stochastic process. The first and second moments of the noise force exerted on particle i, a component of $\boldsymbol{\xi}(t)$, is given by [46]

$$\langle \boldsymbol{\xi_i}(t) \rangle = 0\,, \tag{5.3a}$$

$$\boldsymbol{R_{\xi_{(ij)}}}(\tau) = \langle \boldsymbol{\xi_i}(0), \boldsymbol{\xi_j}(\tau) \rangle = 2\mathrm{k_B T} \boldsymbol{L_{(i)}} \delta_{ij} \delta(\tau)\ . \tag{5.3b}$$

Above, $\langle \cdot \rangle$ denotes an average with respect to the distribution of the realizations of $\boldsymbol{\xi(t)}$. Due to ergodicity, the time and ensemble averages are equal. $\boldsymbol{R_\xi}(\tau) = \mathbb{E}[\boldsymbol{\xi}(0)\boldsymbol{\xi}(\tau)^T]$ is the auto-correlation matrix of the random force $\boldsymbol{\xi}(t)$, of size $3N \times 3N$. The Dirac delta function $\delta(\tau)$ indicates that there is no correlation between different time samples of the random forces. The Kronecker delta function δ_{ij} indicates that random forces acting on different particles are uncorrelated. Therefore, $\boldsymbol{R_\xi}(\tau)$ is a diagonal matrix. i and j represent particle indices, $i, j = 1, 2, \ldots, N$.

5.2.2 Constrained transfer function

The Hessian $\boldsymbol{H}$ is a singular matrix with a rank deficiency of six, due to the degrees of freedom arising from rigid-body rotations and translations of the molecule. Therefore, the transfer function computation introduced in Section 2.3 is ill-conditioned at low frequencies, and not possible at zero frequency. To eliminate the six degrees of freedom, the following set of translational and rotational Eckart's conditions are introduced [91, 92]

$$\sum_{i=1}^{N} m_i\, \Delta \boldsymbol{r} = \boldsymbol{0} \tag{5.4a}$$

$$\sum_{i=1}^{N} m_i\, (\boldsymbol{r^0} \times \Delta \boldsymbol{r}) = \boldsymbol{0}\ , \tag{5.4b}$$

where $\times$ indicates a vector product, $\boldsymbol{r^0} = \boldsymbol{r_i} - \boldsymbol{r_{com}}$ is the displacement from the center of mass, $\boldsymbol{r_{com}}$, and m_i is the mass of the ith particle. Equation 5.4 can be recast in matrix form as follows

$\boldsymbol{C}\Delta\boldsymbol{r} = \boldsymbol{0}$, where

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{M_{(1)}} & \boldsymbol{M_{(2)}} & \dots & \boldsymbol{M_{(N)}} \\ m_1\boldsymbol{R_{(1)}} & m_2\boldsymbol{R_{(2)}} & \dots & m_N\boldsymbol{R_{(N)}} \end{bmatrix} \text{ and } \boldsymbol{R_{(i)}} = \begin{bmatrix} 0 & -\boldsymbol{r^0_{i,z}} & \boldsymbol{r^0_{i,y}} \\ \boldsymbol{r^0_{i,z}} & 0 & -\boldsymbol{r^0_{i,x}} \\ -\boldsymbol{r^0_{i,y}} & \boldsymbol{r^0_{i,x}} & 0 \end{bmatrix} . \tag{5.5}$$

With the incorporation of these constraints, the inversion of a nearly rank-deficient matrix is replaced with the solution of the following constrained least-squares problem

$$\begin{aligned} \underset{\Delta\boldsymbol{r}}{\text{minimize}} \quad & ||\boldsymbol{K}\Delta\boldsymbol{r} - \boldsymbol{F}||^2 \\ \text{subject to} \quad & \boldsymbol{C}\Delta\boldsymbol{r} = \boldsymbol{0} \ . \end{aligned} \tag{5.6}$$

A Lagrange multiplier-based solution of the above problem can be formulated as follows [96]. $\Delta\hat{\boldsymbol{r}}$ is a solution. Lagrangian function with Lagrange multipliers $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \cdots, \lambda_6]^T$:

$$\mathcal{L}(\boldsymbol{\Delta r}, \boldsymbol{\lambda}) = ||\boldsymbol{K}\Delta\boldsymbol{r} - \boldsymbol{F}||^2 + \lambda_1 \boldsymbol{C_1^T}\boldsymbol{\Delta r} + \cdots + \lambda_6 \boldsymbol{C_6^T}\boldsymbol{\Delta r} \ . \tag{5.7}$$

Two optimality conditions are:

$$\begin{aligned} &\frac{\partial\mathcal{L}}{\partial\boldsymbol{\Delta r_i}}(\boldsymbol{\Delta}\hat{\boldsymbol{r}}, \boldsymbol{\lambda}) = 0, \ i = 1, \dots, 3N \ , \\ &\frac{\partial\mathcal{L}}{\partial\boldsymbol{\lambda_i}}(\boldsymbol{\Delta}\hat{\boldsymbol{r}}, \boldsymbol{\lambda}) = 0, \ i = 1, \dots, 6 \ . \end{aligned} \tag{5.8}$$

The second optimality condition yields: $\dfrac{\partial\mathcal{L}}{\partial\boldsymbol{\lambda_i}}(\boldsymbol{\Delta}\hat{\boldsymbol{r}}, \boldsymbol{\lambda}) = \boldsymbol{C_i^T}\Delta\hat{\boldsymbol{r}} = 0$. The first set of conditions gives:

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\Delta r_i}}(\boldsymbol{\Delta}\hat{\boldsymbol{r}}, \boldsymbol{\lambda}) = 2\sum_{j=1}^{3N}(\boldsymbol{K}\boldsymbol{K^T})_{ij}\Delta\hat{\boldsymbol{r}}_{\boldsymbol{(j)}} - 2(\boldsymbol{K^T F})_{\boldsymbol{i}} + \sum_{j=1}^{6}\boldsymbol{\lambda_j}\boldsymbol{C}_i = 0 \ . \tag{5.9}$$

In matrix-vector form

$$\begin{bmatrix} 2\boldsymbol{K^T K} & \boldsymbol{C^T} \\ \boldsymbol{C} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \Delta\hat{\boldsymbol{r}} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} 2\boldsymbol{K^T F} \\ \boldsymbol{0} \end{bmatrix} \ . \tag{5.10}$$

With rearrangements,

$$\begin{bmatrix} \Delta\hat{\boldsymbol{r}} \\ \boldsymbol{\lambda} \end{bmatrix} = \underbrace{\begin{bmatrix} 2\boldsymbol{K}^{\boldsymbol{T}}\boldsymbol{K} & \boldsymbol{C}^{\boldsymbol{T}} \\ \boldsymbol{C} & \boldsymbol{0} \end{bmatrix}^{-1}}_{\begin{bmatrix} \boldsymbol{Z} & \cdot \\ \cdot & \cdot \end{bmatrix}} \begin{bmatrix} 2\boldsymbol{K}^{\boldsymbol{T}}\boldsymbol{F} \\ \boldsymbol{0} \end{bmatrix} . \tag{5.11}$$

The transfer function matrix incorporating the Eckart constraints, denoted by $\boldsymbol{T_c}$, can be readily constructed using the first $3N$ rows and columns of the matrix that is obtained as the inverse of an augmented matrix shown in the RHS of Equation 5.11, denoted by $\boldsymbol{Z}$. Thus,

$$\begin{aligned} \boldsymbol{T_c}(f) &= 2 \cdot \boldsymbol{Z}(f) \cdot \boldsymbol{K}(f)^{\boldsymbol{T}} \\ \Delta\hat{\boldsymbol{r}}(f) &= \boldsymbol{T_c}(f) \cdot \boldsymbol{F}(f) \ . \end{aligned} \tag{5.12}$$

In the rest of the chapter, whenever appropriate, the singularity or ill-conditioning problem in matrix inversion is addressed as above.

5.2.3 Lyapunov formulation

This section presents a Lyapunov-equation based approach in order to show that the mean square fluctuations, i.e., fluctuations under no external force, are independent of the values of the friction coefficients. The rationale behind this discussion is that random noise forces and the viscous friction forces have counteracting effects, as dictated by the Fluctuation-Dissipation theorem [94, 95].

Equation 5.1 can be recast as $6N$ coupled first order differential equations with the state space vector $\boldsymbol{v}$

$$\boldsymbol{v}(t) = \begin{bmatrix} \boldsymbol{\Delta r}(t) \\ \boldsymbol{\Delta \dot{r}}(t) \end{bmatrix} .$$

Under no external force, Equation 5.1 can be expressed as an initial value problem,

with initial conditions $\boldsymbol{v}(0)$, as follows

$$\begin{aligned}\frac{d}{dt}\boldsymbol{v(t)} &= \underbrace{\begin{bmatrix}\mathbf{0} & \boldsymbol{I}\\ -\boldsymbol{M^{-1}H} & -\boldsymbol{M^{-1}L}\end{bmatrix}}_{\boldsymbol{A}}\boldsymbol{v(t)} + \underbrace{\begin{bmatrix}\mathbf{0}\\ \boldsymbol{M^{-1}}\end{bmatrix}}_{\boldsymbol{B}}\boldsymbol{\xi}(t)\\ \boldsymbol{v}(0) &= \begin{bmatrix}\boldsymbol{\Delta r}(0)\\ \boldsymbol{\Delta \dot{r}}(0)\end{bmatrix}\ .\end{aligned} \tag{5.13}$$

Due to the fact that white noise $\boldsymbol{\xi}(t)$ can be expressed as the formal time derivative of a Wiener Process, $\boldsymbol{W}$, Equation 5.13 is written in differential form as follows

$$d\boldsymbol{v} = \boldsymbol{A}\boldsymbol{v}\,dt + \boldsymbol{B}\,d\boldsymbol{W}\ . \tag{5.14}$$

Noise characteristics are as follows:

$$\mathbb{E}[\boldsymbol{\xi}(t)] = \mathbf{0} \tag{5.15a}$$

$$\mathbb{E}[\boldsymbol{\xi}(t)\,\boldsymbol{\xi}^T(t+\tau)] = 2\,\mathrm{k_B T}\,\boldsymbol{L}\,\delta(\tau)\ , \tag{5.15b}$$

where $\mathbb{E}[\cdot]$ denotes the probabilistic expectation operator. Taking the expectation of both sides of Equation 5.14, and using the property in Equation 5.15a

$$\mathbb{E}[d\boldsymbol{v}] = \boldsymbol{A}\,\mathbb{E}[\boldsymbol{v}]\,dt\ . \tag{5.16}$$

The zero time-lag correlation matrix is defined as

$$\boldsymbol{R_{vv}} = \mathbb{E}[\boldsymbol{v}\,\boldsymbol{v}^T]\ . \tag{5.17}$$

From Equations 5.14 and 5.17, and using Ito's product rule [97] (from stochastic Ito calculus), the following equation is derived

$$\begin{aligned}d(\mathbb{E}[\boldsymbol{v}\,\boldsymbol{v^T}]) =& \mathbb{E}[\boldsymbol{dv}\,\boldsymbol{v^T}] + \mathbb{E}[\boldsymbol{v}\,\boldsymbol{dv^T}] + \mathbb{E}[\boldsymbol{dv}\,\boldsymbol{dv^T}]\\ =& \boldsymbol{A}\,\mathbb{E}[\boldsymbol{v}\,\boldsymbol{v^T}]dt + \boldsymbol{B}\,\mathbb{E}[\boldsymbol{dW}\,\boldsymbol{v^T}] + \mathbb{E}[\boldsymbol{v}\,\boldsymbol{v^T}]\boldsymbol{A^T}dt + \mathbb{E}[\boldsymbol{v}\,\boldsymbol{dW^T}]\boldsymbol{B^T} +\\ & \mathbb{E}[\boldsymbol{dv}\,\boldsymbol{dv^T}]\ .\end{aligned} \tag{5.18}$$

The second and fourth terms in Equation 5.18 are zero due to the fact that $\boldsymbol{v}$ and $\boldsymbol{dW}$ are uncorrelated [27], i.e.,

$$\begin{aligned}\mathbb{E}[\boldsymbol{v}\,\boldsymbol{dW^T}] &= 0\\ \mathbb{E}[\boldsymbol{dW}\,\boldsymbol{v^T}] &= 0\ .\end{aligned} \tag{5.19}$$

The additional term in the stochastic calculus product rule in Equation 5.18 can be expanded as

$$\begin{aligned}\mathbb{E}[\boldsymbol{dv}\,\boldsymbol{dv}^{\boldsymbol{T}}] =& \mathbb{E}[(\boldsymbol{A}\boldsymbol{v}dt + \boldsymbol{B}\boldsymbol{dW})(\boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{A}^{\boldsymbol{T}}dt + \boldsymbol{dW}^{\boldsymbol{T}}\boldsymbol{B}^{\boldsymbol{T}})] \\ =& \boldsymbol{A}\,\mathbb{E}[\boldsymbol{v}\boldsymbol{v}^{\boldsymbol{T}}]\boldsymbol{A}^{\boldsymbol{T}}dt^2 + \boldsymbol{A}\,\mathbb{E}[\boldsymbol{v}\boldsymbol{dW}^{\boldsymbol{T}}]\boldsymbol{B}^{\boldsymbol{T}}dt + \\ & \boldsymbol{B}\,\mathbb{E}[\boldsymbol{dW}\boldsymbol{v}^{\boldsymbol{T}}]\boldsymbol{A}^{\boldsymbol{T}}dt + \boldsymbol{B}\,\mathbb{E}[\boldsymbol{dW}\,\boldsymbol{dW}^{\boldsymbol{T}}]\boldsymbol{B}^{\boldsymbol{T}}\ .\end{aligned} \tag{5.20}$$

The first three terms in Equation 5.20 are set to zero due to the following that follow from rules of stochastic Ito calculus (The second and third terms vanish, also due to Equation 5.19.)

$$dt^2 \approx 0 \quad dW\,dt \approx 0\ . \tag{5.21}$$

From the noise properties in Equation 5.15b

$$\mathbb{E}[\boldsymbol{dW}\,\boldsymbol{dW}^{\boldsymbol{T}}] = 2\mathrm{k_B T}\boldsymbol{L}\,dt\ . \tag{5.22}$$

Then, Equation 5.18 becomes

$$d\boldsymbol{R}_{\boldsymbol{vv}} = \boldsymbol{A}\boldsymbol{R}_{\boldsymbol{vv}}dt + \boldsymbol{R}_{\boldsymbol{vv}}\boldsymbol{A}^{\boldsymbol{T}}dt + 2\mathrm{k_B T}\boldsymbol{B}\boldsymbol{L}\boldsymbol{B}^{\boldsymbol{T}}dt\ . \tag{5.23}$$

Zero time-lag covariance matrix $\boldsymbol{C}_{\boldsymbol{vv}}$ is defined as

$$\boldsymbol{C}_{\boldsymbol{vv}} = \mathbb{E}[(\boldsymbol{v} - \mathbb{E}[\boldsymbol{v}])(\boldsymbol{v} - \mathbb{E}[\boldsymbol{v}])^T] = \boldsymbol{R}_{\boldsymbol{vv}} - \mathbb{E}[\boldsymbol{v}]\,\mathbb{E}[\boldsymbol{v}^{\boldsymbol{T}}]\ . \tag{5.24}$$

Differential form of $\boldsymbol{C}_{\boldsymbol{vv}}$:

$$d\boldsymbol{C}_{\boldsymbol{vv}}(t) = d\boldsymbol{R}_{\boldsymbol{vv}} - \mathbb{E}[d\boldsymbol{v}]\ \mathbb{E}[\boldsymbol{v}^{\boldsymbol{T}}] - \mathbb{E}[\boldsymbol{v}]\ \mathbb{E}[d\boldsymbol{v}^{\boldsymbol{T}}] - \mathbb{E}[d\boldsymbol{v}]\mathbb{E}[d\boldsymbol{v}^{\boldsymbol{T}}]\ . \tag{5.25}$$

Utilizing Equation 5.16 and expanding $d\boldsymbol{R}_{\boldsymbol{vv}}$ according to Equation 5.23 yields

$$d\boldsymbol{C}_{\boldsymbol{vv}}(t) = \boldsymbol{A}\boldsymbol{C}_{\boldsymbol{vv}}dt + \boldsymbol{C}_{\boldsymbol{vv}}\boldsymbol{A}^{\boldsymbol{T}}dt + 2\mathrm{k_B T}\boldsymbol{B}\boldsymbol{L}\boldsymbol{B}^{\boldsymbol{T}}dt\ . \tag{5.26}$$

In the limit as time goes to infinity, $\boldsymbol{v}(t)$ approaches to a (wide-sense) stationary process where $\mathbb{E}[\boldsymbol{v}]$ and $\boldsymbol{C}_{\boldsymbol{vv}}$ become independent of time and we obtain Equation 5.27b, a *Lyapunov equation* [98].

$$\lim_{t\to\infty} \mathbb{E}[\boldsymbol{v}] = \boldsymbol{0} \tag{5.27a}$$

$$\lim_{t\to\infty} \boldsymbol{A}\boldsymbol{C}_{\boldsymbol{vv}} + \boldsymbol{C}_{\boldsymbol{vv}}\boldsymbol{A}^{\boldsymbol{T}} + 2\mathrm{k_B T}\boldsymbol{B}\boldsymbol{L}\boldsymbol{B}^{\boldsymbol{T}} = \boldsymbol{0}\ . \tag{5.27b}$$

We note that, due to Equation 5.27a, $\boldsymbol{C_{vv}} = \boldsymbol{R_{vv}}$, however we stick to the covariance notation. Using the symmetricity of covariance matrices, $\boldsymbol{C_{vv}}$ is structured in block form as

$$\begin{aligned} \boldsymbol{C_{vv}} &= \begin{bmatrix} \mathbb{E}[\boldsymbol{\Delta r \Delta r^T}] & \mathbb{E}[\boldsymbol{\Delta r \Delta \dot{r}^T}] \\ \mathbb{E}[\boldsymbol{\Delta \dot{r} \Delta r^T}] & \mathbb{E}[\boldsymbol{\Delta \dot{r} \Delta \dot{r}^T}] \end{bmatrix} \\ \boldsymbol{C_{vv}} &= \begin{bmatrix} \boldsymbol{C_{00}} & \boldsymbol{C_{01}} \\ \boldsymbol{C_{01}}^{\boldsymbol{T}} & \boldsymbol{C_{11}} \end{bmatrix} . \end{aligned} \tag{5.28}$$

Inserting $\boldsymbol{A}$ defined in Equation 5.13 and the block form of $\boldsymbol{C_{vv}}$, with $\boldsymbol{H} = \boldsymbol{H^T}$, Equation 5.27b expands as

$$\begin{aligned} &\begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} \\ -\boldsymbol{M^{-1}H} & -\boldsymbol{M^{-1}L} \end{bmatrix} \begin{bmatrix} \boldsymbol{C_{00}} & \boldsymbol{C_{01}} \\ \boldsymbol{C_{01}}^{\boldsymbol{T}} & \boldsymbol{C_{11}} \end{bmatrix} + \begin{bmatrix} \boldsymbol{C_{00}} & \boldsymbol{C_{01}} \\ \boldsymbol{C_{01}}^{\boldsymbol{T}} & \boldsymbol{C_{11}} \end{bmatrix} \begin{bmatrix} \boldsymbol{0} & -\boldsymbol{HM^{-1}} \\ \boldsymbol{I} & -\boldsymbol{M^{-1}L} \end{bmatrix} + \\ &\qquad 2\mathrm{k_B T} \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M^{-2}L} \end{bmatrix} = \boldsymbol{0} , \end{aligned} \tag{5.29}$$

which can be further manipulated as

$$\begin{aligned} &\begin{bmatrix} \boldsymbol{C_{01}}^{\boldsymbol{T}} & \boldsymbol{C_{11}} \\ -\boldsymbol{M^{-1}HC_{00}} - \boldsymbol{M^{-1}L}\boldsymbol{C_{01}}^{\boldsymbol{T}} & -\boldsymbol{M^{-1}HC_{01}} - \boldsymbol{M^{-1}LC_{11}} \end{bmatrix} + \\ &\begin{bmatrix} \boldsymbol{C_{01}} & -\boldsymbol{C_{00}HM^{-1}} - \boldsymbol{C_{01}M^{-1}L} \\ \boldsymbol{C_{11}} & -\boldsymbol{C_{01}}^{\boldsymbol{T}}\boldsymbol{HM^{-1}} - \boldsymbol{C_{11}M^{-1}L} \end{bmatrix} + 2\mathrm{k_B T} \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M^{-2}L} \end{bmatrix} = \boldsymbol{0} . \end{aligned} \tag{5.30}$$

Equation 5.30 leads to the following set of equations

$$\boldsymbol{C_{01}}^{\boldsymbol{T}} + \boldsymbol{C_{01}} = \boldsymbol{0} \tag{5.31a}$$

$$\boldsymbol{C_{11}} - \boldsymbol{C_{00}HM^{-1}} - \boldsymbol{C_{01}M^{-1}L} = \boldsymbol{0} \tag{5.31b}$$

$$-\boldsymbol{M^{-1}HC_{00}} - \boldsymbol{M^{-1}L}\boldsymbol{C_{01}}^{\boldsymbol{T}} + \boldsymbol{C_{11}} = \boldsymbol{0} \tag{5.31c}$$

$$-\boldsymbol{M^{-1}HC_{01}} - \boldsymbol{M^{-1}LC_{11}} - \boldsymbol{C_{01}}^{\boldsymbol{T}}\boldsymbol{HM^{-1}} - \boldsymbol{C_{11}M^{-1}L} + 2\mathrm{k_B T}\boldsymbol{M^{-2}L} = \boldsymbol{0} . \tag{5.31d}$$

Equation 5.31a implies that $\boldsymbol{C_{01}} = -\boldsymbol{C_{01}}^{\boldsymbol{T}}$, i.e., $\boldsymbol{C_{01}}$ is antisymmetric. Then, the

above set of equations yield

$$\boldsymbol{C_{11} - C_{00} H M^{-1} - C_{01} M^{-1} L = 0} \tag{5.32a}$$

$$\boldsymbol{- M^{-1} H C_{00} + M^{-1} L C_{01} + C_{11} = 0} \tag{5.32b}$$

$$\boldsymbol{- M^{-1} H C_{01} - M^{-1} L C_{11} + C_{01} H M^{-1} - C_{11} M^{-1} L + 2 k_B T M^{-2} L = 0} \ . \tag{5.32c}$$

From Equations 5.32a and 5.32b

$$\boldsymbol{C_{11} = C_{00} H M^{-1} + C_{01} M^{-1} L} \tag{5.33a}$$

$$\boldsymbol{= M^{-1} H C_{00} - M^{-1} L C_{01}} \ . \tag{5.33b}$$

Noting the symmetricity of covariance matrices: $\boldsymbol{C_{00} = C_{00}^{T}}$ and $\boldsymbol{C_{11} = C_{11}^{T}}$. Then, inserting Equation 5.33 into Equation 5.32c gives

$$\begin{aligned} &\boldsymbol{- M^{-1} H C_{01} + C_{01} H M^{-1} - M^{-1} L (C_{00} H M^{-1} + C_{01} M^{-1} L)} \\ &\boldsymbol{- (M^{-1} H C_{00} - M^{-1} L C_{01}) M^{-1} L + 2 k_B T M^{-2} L = 0} \ . \end{aligned} \tag{5.34}$$

Multiplying the above by $\boldsymbol{M}$ both from the left and right results in

$$\boldsymbol{- H C_{01} M + M C_{01} H - L C_{00} H - H C_{00} L + 2 k_B T L = 0} \ . \tag{5.35}$$

When $\boldsymbol{C_{00}} = k_B T \boldsymbol{H^{-1}}$, Equations 5.33a and 5.33b together read

$$\boldsymbol{C_{01} M^{-1} L + M^{-1} L C_{01} = 0} \ , \tag{5.36}$$

implying that $\boldsymbol{C_{01} = 0}$. We then observe that $\boldsymbol{C_{00}} = k_B T \boldsymbol{H^{-1}}$ satisfies Equation 5.35, based on the fact that the first two terms are equal to zero with this form of solution for $\boldsymbol{C_{00}}$, as shown above. We also note that $\boldsymbol{M}$, $\boldsymbol{M^{-1}}$, $\boldsymbol{L}$ are all diagonal matrices with positive-valued entries. This shows that the mean square displacements are independent of the chosen $\boldsymbol{L}$. We note also that $\boldsymbol{H}$ is singular, thereby its inverse is calculated either as a pseudo-inverse, or alternatively, by adapting the constraint-based solution that is introduced in Section 5.2.2. The frequency-dependent terms are zero in the transfer function ($\boldsymbol{T}$) definition at zero frequency in Equation 2.11, i.e., $\boldsymbol{T(0) = H^{-1}}$. To address the singularity problem, rotational

and translational constraints are introduced into the transfer function, leading to Equation 5.12. Then, the constrained transfer function at zero frequency, $\boldsymbol{T_c(0)}$, denoted as $\boldsymbol{{H_c}^{-1}}$, can be calculated as

$$\boldsymbol{{H_c}^{-1} = T_c(0) = 2 \cdot Z(0) \cdot H^T} \text{ with} \tag{5.37a}$$

$$\begin{bmatrix} \boldsymbol{Z(0)^{3N \times 3N}} & \cdot \\ \cdot & \cdot \end{bmatrix} = \begin{bmatrix} \boldsymbol{2H^T H} & \boldsymbol{C^T} \\ \boldsymbol{C} & \boldsymbol{0} \end{bmatrix}^{-1}, \tag{5.37b}$$

where $\boldsymbol{C}$ corresponds to the constraints matrix, defined in Equation 5.5.

As an alternative approach in order to demonstrate that the mean square displacements are independent of the friction values, Figure 5.1 shows the numerical solution of Equation 5.27b for $\boldsymbol{C_{00}}$. Due to the previously mentioned invertibility issues of $\boldsymbol{H}$, Lyapunov equation does not have a unique solution with matrix $\boldsymbol{A}$ arising from a rank-deficient $\boldsymbol{H}$. In order to overcome this problem in the numerical solution, $\boldsymbol{H_c}$ is utilized instead. Figure 5.1 is drawn for various $\boldsymbol{L}$ values. The same figure is obtained in all trials. Therefore, it can be concluded that due to the link between thermal noise and viscous friction, residue fluctuations are independent of the friction values.

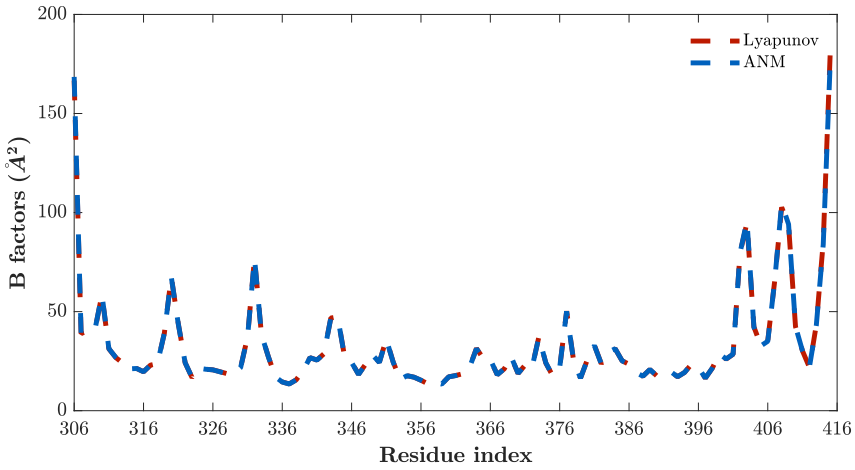


Figure 5.1: Numerical Solution of Lyapunov equation

5.2.4 Power spectral density (PSD)

The power spectral density describes the content of a signal in a frequency decomposed manner. It can be computed either directly from the squared amplitudes of the frequency domain representation (for deterministic signals), as in Equation 5.38a below, or as the Fourier transform of the auto-correlation function (for stochastic signals), as in Equation 5.38b, due to the Wiener-Khinchin theorem [99]. $\boldsymbol{S_X}$ is the PSD of a signal X

$$\boldsymbol{S_X}(f) = |\boldsymbol{\mathcal{F}}[X(t)]|^2 \tag{5.38a}$$

$$= \boldsymbol{\mathcal{F}}[\boldsymbol{R_X}(\tau)] \ , \tag{5.38b}$$

where $\boldsymbol{\mathcal{F}}$ denotes the Fourier transform, defined with $\int_{-\infty}^{\infty} \mathrm{d}\tau X(\tau) \exp(-j2\pi f\tau)$ for the signal X. Equation 5.38a is used for external force excitations that are deterministic, Equation 5.38b is used for noise sources and stochastic fluctuations.

5.2.4.1 PSD of noise sources

The PSD of the noise force acting on the i^{th} particle can be determined by inserting the autocorrelation $\boldsymbol{R_{\xi_{(ij)}}}(\tau)$ defined in Equation 5.3b into Equation 5.38b

$$\boldsymbol{S_{\xi_{(i)}}} = 2\mathrm{k_B T}\boldsymbol{L_{(i)}} \ . \tag{5.39}$$

The above PSD is constant over the frequency spectrum, corresponding to *white noise*.

5.2.4.2 PSD of external force excitations

A deterministic external force that acts on certain nodes mimicks the ligand binding event. Although the proposed method does not impose any restriction on the form of the external force $\boldsymbol{F_e}$, we consider it as a dynamic, sinusoidal perturbation whose frequency is swept over a certain range, allowing us to quantify the effect of perturbation frequency. The external force applied to particle j is in the form

$$\boldsymbol{F_{e(j)}} = \boldsymbol{F_j} \cos(2\pi f_c t) \ ,$$

where $\boldsymbol{F_j}$ is a 3×1 vector that contains the force components along the coordinate axes applied on the jth particle, and f_c is the frequency.

The PSD of this sinusoidal force at the input site is given by

$$\boldsymbol{S_{F_e(j)}}(f) = \begin{bmatrix} \boldsymbol{F_{j_x}^2}/4 & 0 & 0 \\ 0 & \boldsymbol{F_{j_x}^2}/4 & 0 \\ 0 & 0 & \boldsymbol{F_{j_z}^2}/4 \end{bmatrix} \delta(f - f_c) \ . \tag{5.40}$$

5.2.5 From excitation PSDs to displacement PSDs

For multi-input multi-output LTI systems, the cross spectral density matrix of the outputs, $\boldsymbol{S_Y}$, can be calculated as follows by using the cross spectral density matrix of the inputs, $\boldsymbol{S_X}$, and the transfer function $\boldsymbol{T}$ (please see [100] for the derivation)

$$\boldsymbol{S_Y}(f) = \boldsymbol{T}(f)\, \boldsymbol{S_X}(f)\, \boldsymbol{T}(f)^* \ , \tag{5.41}$$

where $\cdot^*$ denotes complex conjugate. When the input is the force excitation (either due to noise or external force), and the output is the fluctuations/displacements of node positions, the above relationship can be written using the constrained transfer function defined in Equation 5.12 as

$$\boldsymbol{S_{\Delta r}^F}(f) = \boldsymbol{T_c}(f)\, \boldsymbol{S_F}(f)\, \boldsymbol{T_c}(f)^* \ . \tag{5.42}$$

We separately compute the PSDs of the output due to noise and deterministic excitations by first setting $\boldsymbol{F} = \boldsymbol{\xi}$ and then $\boldsymbol{F} = \boldsymbol{F_e}$. We note that $\boldsymbol{S_{\Delta r}^F}$ is a full matrix. The diagonal entries represent the spectral densities, whereas the off-diagonal entries are the cross-spectral densities. We focus on the diagonal entries in this study. Hence, we introduce the following notation

$$\boldsymbol{D}(\boldsymbol{S_{\Delta r}^F}) = (\boldsymbol{D}(\boldsymbol{S_{\Delta r(1)}^F}), \boldsymbol{D}(\boldsymbol{S_{\Delta r(2)}^F}), \cdots, \boldsymbol{D}(\boldsymbol{S_{\Delta r(N)}^F}))$$

for the $3N \times 1$ vector formed from the diagonal elements of $\boldsymbol{S_{\Delta r}^F}$.

5.2.6 Characterization of equilibrium fluctuations based on the PSDs of displacements due to noise

The partition function of the system, whose potential energy profile is given in Equation 2.5, can be calculated over all possible displacements with

$Z = \int \mathrm{d}\Delta r \exp\left(-\beta \Delta \boldsymbol{r}^T \boldsymbol{H}(\bar{\boldsymbol{r}}) \Delta \boldsymbol{r}\right)$, where $\beta = 1/\mathrm{k_B}T$, $\mathrm{k_B}$ is the Boltzmann's constant, and T is the absolute temperature. The correlations of the fluctuations between sites i and j can be obtained from this partition function as [101]

$$\langle \Delta \boldsymbol{r_i} \cdot \Delta \boldsymbol{r_j} \rangle = \frac{1}{Z} \int \mathrm{d}\Delta r (\Delta \boldsymbol{r_i} \cdot \Delta \boldsymbol{r_j}) \exp(-\beta \Delta \boldsymbol{r}^T \boldsymbol{H} \Delta \boldsymbol{r}) \ , \tag{5.43}$$

which can be simplified to

$$\langle \Delta \boldsymbol{r_i} \cdot \Delta \boldsymbol{r_j} \rangle = \mathrm{tr}(\boldsymbol{H}^{-1}_{(ij)})/\beta \ . \tag{5.44}$$

Thus, mean square fluctuations for particle i is

$$\langle \Delta \boldsymbol{r_i}^2 \rangle = \mathrm{tr}(\boldsymbol{H}^{-1}_{(ii)})/\beta \ , \tag{5.45}$$

where $\boldsymbol{H}_{(ij)}$ corresponds to the 3×3 submatrix for particles i and j, and tr denotes the trace. Since the Hessian is singular as discussed before, its pseudo-inverse ($\boldsymbol{H}^+$) is computed through the use of the non-zero eigenvalues, $\boldsymbol{\mu}_k$, and the corresponding eigenvectors, $\boldsymbol{u}_k$, as follows

$$\boldsymbol{H}^+ = \sum_{k=1}^{3N-6} \frac{1}{\mu_k} \boldsymbol{u_k} \boldsymbol{u_k}^T \ . \tag{5.46}$$

The mean square fluctuations computed as such are proportional to the B factors determined from X-ray crystallography, also known as temperature factors [93], denoted by B:

$$B \propto \langle \Delta \boldsymbol{r}^2 \rangle \ . \tag{5.47}$$

The total power in a WSS stochastic signal X can be expressed in terms of the variance of the signal, i.e., zero time-lagged auto-correlation, which is also given by the integral of its PSD over all frequencies, as below

$$P_X = \langle X(t)^2 \rangle = R_X(0) = \int_{-\infty}^{\infty} \mathrm{d}f \ S_X(f) \ . \tag{5.48}$$

In the absence of external force excitations, the mean square fluctuation of node i can be thus computed as the sum of integrals of the displacement PSDs along the coordinate axes as

$$\langle \Delta \boldsymbol{r_i}^2 \rangle = 2 \sum_{\mathrm{k=x,y,z}} \int_0^{\infty} \mathrm{d}f \ (\boldsymbol{S}_{\Delta \boldsymbol{r}}{}^{\boldsymbol{\xi}})_{i,\mathrm{k}}(f) \ , \tag{5.49}$$

where the factor 2 is due to the symmetry of the two-sided PSD for negative and positive frequencies. This PSD-based technique thereby offers a previously overlooked route to computing mean square fluctuations. Previously published studies mostly focused on the frequency distribution of the B factors at a set of discrete frequency points, known as the normal mode frequencies. The PSD-based approach allows the frequency decomposed analysis of B factors over any continuous frequency range of interest.

5.2.7 Signal-to-Noise ratio (SNR) and channel capacity

We define SNR as the ratio of the PSD of the displacements measured at the output due to external force excitations alone and the displacements due to only noise:

$$\mathrm{SNR}(f) = \frac{\boldsymbol{S}^{\boldsymbol{F_e}}_{\boldsymbol{\Delta r}}(f)}{\boldsymbol{S}^{\boldsymbol{\xi}}_{\boldsymbol{\Delta r}}(f)} \ . \tag{5.50}$$

According to the integral-form of the Shannon-Hartley theorem [90], the channel capacity that can be attained over a noisy communication channel is defined as

$$C = \int_{BW} \mathrm{d}f \log_2 \left(1 + \mathrm{SNR}(f)\right) \ , \tag{5.51}$$

where the integral is computed over a frequency band BW of interest. We choose $[0, \ \infty)$ as the frequency interval. Then, the channel capacity between the input node j and output node i can be calculated as the sum of the channel capacities along the coordinates axes

$$C_{ij} = \sum_{k=x,y,z} \int_0^\infty \mathrm{d}f \log_2 \left(1 + \frac{\boldsymbol{D}(\boldsymbol{S}^{\boldsymbol{F_{e(j)}}}_{\boldsymbol{\Delta r\,(i),k}})}{\boldsymbol{D}(\boldsymbol{S}^{\boldsymbol{\xi}}_{\boldsymbol{\Delta r\,(i),k}})}\right) \ , \tag{5.52}$$

where F_{ej} indicates that the deterministic force excitation is applied to node j only but results in dynamic displacements at all the others. In computing $\boldsymbol{S}^{\boldsymbol{F_{e(j)}}}_{\boldsymbol{\Delta r\,(i),k}}$, the strength of the input excitation is kept constant while its frequency is swept. The strength of the resulting response will vary not only as a function of the output node but also over the frequency range, as determined by the internal network dynamics. Thus, while information (signal) transmission between two particular nodes may be inefficient at a certain frequency, it may be enhanced at another. On the other

hand, random noise excites all nodes simultaneously, at all frequencies with equal strength, and it is shaped by the network dynamics resulting in fluctuations at all nodes with a colored frequency composition as opposed to the white noise excitation. The extent of fluctuations due to noise is also a function of the output node. We emphasize here the distinction between signal transmission and noise propagation through the network. While the signal permeates throughout the network from a single entry point, noise enters from everywhere. Thus, the frequency decomposed profiles of node displacements due to either the signal or the noise alone may be quite different, resulting in a colored frequency composition for SNR with possibly non-monotonic behavior.

5.3 Methods

Figure 5.2 presents a pictorial overview of the methods described in Section 5.2. The protein is represented as a network, and the protein-solution interactions are modeled as random noise forces that act on every node. In contrast, the effect of the ligand is captured by a deterministic, oscillatory external force, which is applied to a single node (or simultaneously to a set of nodes), denoted as the input node(s). The goal is to examine how the effect of the external force excitation is transmitted to the output node in the presence of all of the noise sources. The random noise is white, i.e., has a constant PSD over the frequency interval of interest. The magnitude of the external force is held constant as a function of frequency. The PSDs measured at the output node are calculated using the input PSDs and the frequency-dependent transfer functions, quantifying the displacements in response to both deterministic and noise excitations. We underline that although noise forces are applied to all of the nodes, noise PSD only for the input node, upon which the external force is applied, is shown in the pictorial overview for brevity. The methods proposed in this thesis enable a frequency decomposed analysis, by sweeping the perturbation frequency in a frequency range of interest. The gray bars in the PSDs represent frequency sweeping, while the highlighted bars correspond to the perturbation frequency that is currently under investigation. The PSD plots at the output show the frequency-

dependent characteristics. SNR is calculated using the output PSDs, by simply taking the ratio of the PSDs due to deterministic and noise excitations at each frequency. The channel capacity is computed via an integral of the SNR over a frequency band. In Figure 5.2, the color green represents noise and red is used for the external force. Yellow and purple are used to distinguish the input and output nodes, respectively.

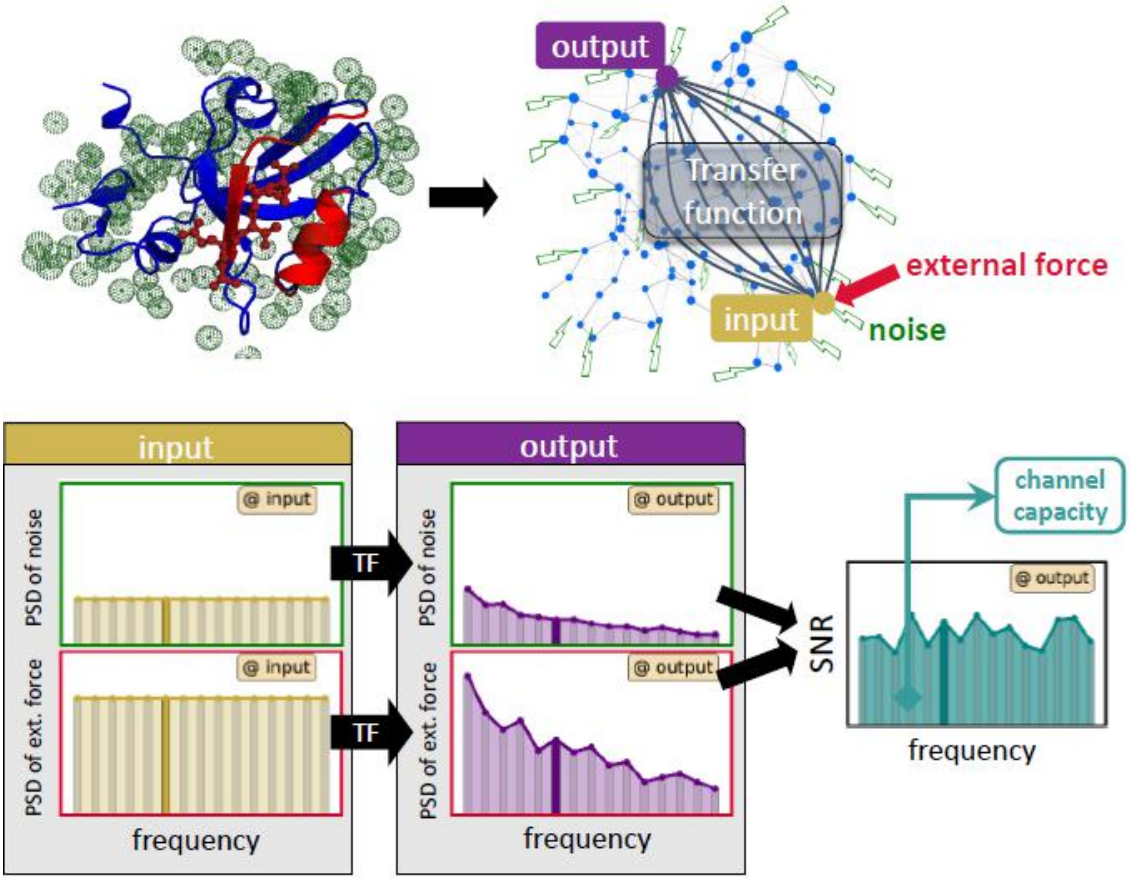


Figure 5.2: A simplified pictorial overview of the work flow. The signal permeates throughout the network from a single entry point, the input, and converges at the output. Noise enters the network from everywhere and every noise component has an impact at the output. Although the noise forces act on every node, only some of them are shown in the figure to reduce clutter. The network was drawn using the NAPS web-server provided in [1].

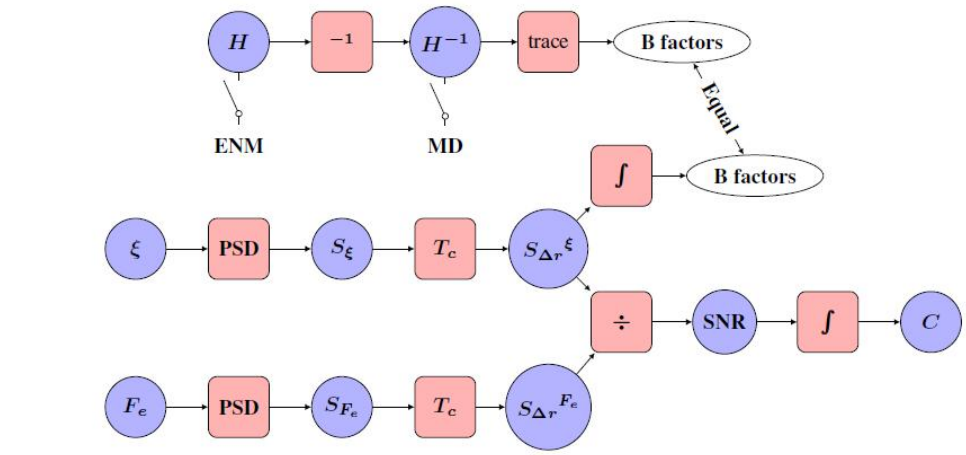


Figure 5.3: Schematic of the work flow.

An algorithmic summary of the proposed methods is presented in Figure 5.3. The top flow line schematically shows the steps for obtaining the B factors from the Hessian. We use two alternate methods for Hessian construction, described in Sections 3.3 and 3.2.1. One approach relies on molecular dynamics (MD) ($\boldsymbol{H}^{\mathrm{MD}}$) while the other is based on ENMs ($\boldsymbol{H}^{\mathrm{ANM}}$). In Figure 5.3, the two parallel flow lines below are for the operations involved in obtaining the output displacement PSDs from the input PSDs, including the computation of the constrained transfer functions, both for deterministic excitations and noise. As an alternative to direct inversion of the Hessian, the PSDs of the displacements due to noise not only lead to the B factors, but also enable the frequency decomposed analysis of the equilibrium fluctuations. SNRs are computed as the ratio of the displacement PSDs due to external force excitations and to those due to noise, subsequently integrated to obtain the channel capacities.

5.3.1 Molecular dynamics simulation details

We considered two distinct systems for simulations, the PDZ3 domain with (PDBID:1be9) and without (PDBID:1bfe) the associated ligand. Starting from the PDB structures, missing atoms were added to the structure using the PyMOL mutagenesis tool [2]. Each structure was then solvated with sufficient number of water molecules to "fill" a cubic box with sides of length 6.6 nm (8804 and 8845 water molecules for the system with and without the ligand, respectively), using the *gmx solvate* command in the GROMACS package [102]. Additionally, the minimum number of ions were added to neutralize the system (2 Na ions for the system with the ligand and 1 Na ion for the system without the ligand). The Amber99SB-ILDN force field [103, 104] was used to model the protein interactions, while the TIP3P model [105] was used to model the water interactions.

All simulations were performed with the Gromacs 5.0.8 simulation suite [102] with a 2 fs integration time step, while using the LINCS [106] algorithm to rigidly constrain all bonds that involve H atoms. All simulations employed the Gromacs leap-frog integrator [107] and periodic boundary conditions [108]. Electrostatic interactions were treated with the particle mesh Ewald method [109], using the default Gromacs parameters. Short-ranged van der Waals interactions and also the real space contribution to the electrostatic interactions were truncated at 1.2 nm in all simulations. After energy minimization, each system was annealed from 0 to 300 K over a 500 ps time frame, followed by a 5 ns equilibration simulation in the NVT ensemble using the Berendsen thermostat [110] with a temperature coupling constant of 0.5. Subsequently, an NPT equilibration was performed using the Berendsen thermostat and barostat with coupling coefficients of 0.1 for both and a compressibility of 4.5×10^{-5} bar^{-1} (corresponding to the compressibility of water at 1 atm and 300 K). The average volume from this simulation (corresponding to box side lengths of 6.56797 nm and 6.56372 nm for the systems with and without the ligand, respectively) was used to select an initial structure for the production simulation. Production simulations were performed for 100 ns in the NPT ensemble, using the velocity rescaling thermostat [111] with a temperature coupling coefficient of 0.1 and

the Parrinello-Rahmen barostat with a pressure coupling coefficient of 2.0. Configurations, velocities, and forces were saved every 2 ps for subsequent analysis. We characterized the local impact of the ligand by calculating the average force that all atoms of the ligand exert on each of the C_α atoms of the binding pocket residues (residues $[320 - 328]$ and $[371 - 380]$).

5.3.2 External force excitations

Figure 5.4 presents a schematic view of the two different techniques we used in order to probe allosteric behavior, namely, *per residue scan* and *binding pocket (BP) excitation*. An external force is applied to only one residue in the first case, whereas multiple residues are perturbed simultaneously in the latter. With *per residue scan*, the goal is to identify the residue pairs that are likely to interact allosterically. Each residue is separately perturbed with a sinusoidal force whose frequency is swept from 0 to the maximum of the normal mode frequencies. The force in this case has a unit magnitude but numerous directions are sampled on a spherical grid. The channel capacities from the input to the rest of the residues are computed for each force direction, and a distinct direction for every output residue that maximizes its capacity is identified. *BP excitation* models ligand binding to the protein in a more realistic manner by simultaneously exciting multiple residues that are previously known to be in a certain binding pocket. The responses of all of the remaining residues are monitored. Instead of sampling the force directions on a grid, the relative amplitudes and the directions of the external forces that act on the binding pocket residues are directly obtained from MD simulations as described in detail in Section 5.3.1, while the sinusoidal form and the frequency range are as in *per residue scan*. The channel capacities calculated as such are normalized (in a min-max sense such that the range of the channel capacity values from the minimum to the maximum are linearly transformed into the range 0 to 1) in order to attain a distribution of capacities within a protein. In the normalization process, the values above a predefined threshold are set to 1, while the remaining values are scaled to be in the interval $[0, 1]$.

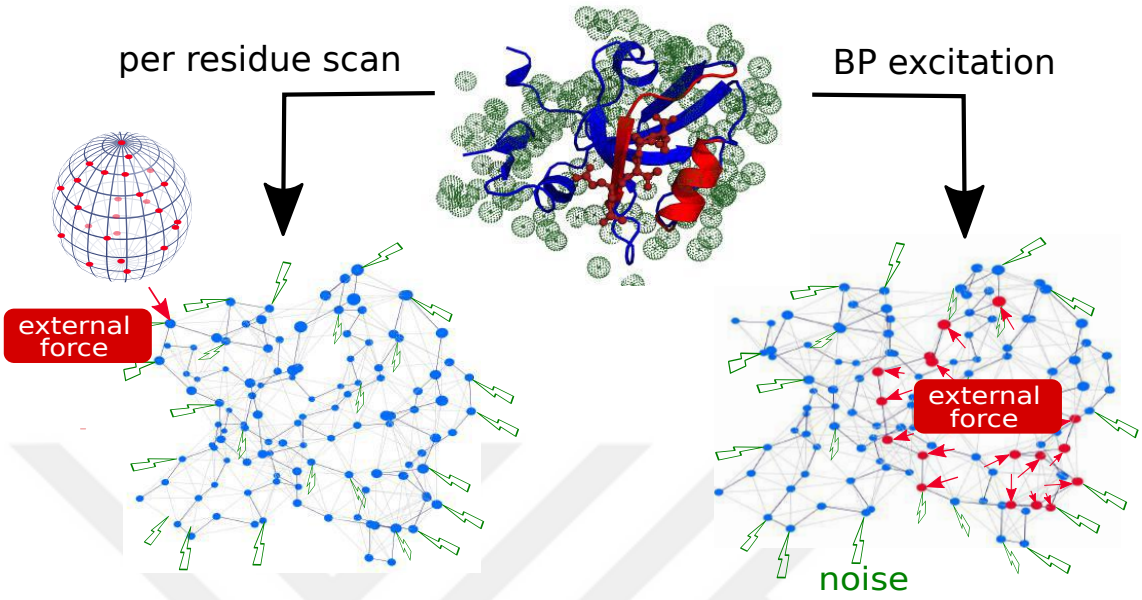


Figure 5.4: Schematic view of the two different techniques used in order to probe allosteric behavior.

5.3.3 Parameters

Residues are represented by their C_α atoms. A 200-point Gauss-Legendre quadrature scheme from NumPy package is used to numerically calculate the integrals [112, 113]. η, which is utilized in the friction coefficient calculation in Equation 5.2 and subsequently in Equation 5.1, is taken as the dynamic viscosity of water at 310 K, that is equal to $6.7807 \cdot 10^{-4}$ Pa $\cdot$ s. Masses and volumes of the residues are retrieved from [114] and [115], respectively, also listed in Table 8.2. However, as shown in Section 5.2.3, the mean square fluctuations are independent of the friction coefficient values due to the link between noise PSDs and friction in the formulation in Equation 5.1. We utilized ANMs with a cut-off distance of 15 Å and the same spring constant for all interactions. In *per residue scan*, the force directions are sampled from a 25 point equally-spaced spherical grid. The values above 90% of the maximum channel capacity value are normalized to a value of 1. In the *BP excitation*, the force exerted by the ligand on the binding pocket residues are determined from MD simulations of the holo form as described in Section 5.3.1. The rotation

matrix calculated for the superposition of the apo and holo forms is then applied to this external force vector so that it can be used as a force excitation for the apo form.

5.4 Results

We demonstrate the utility of the proposed methods on a well-studied PDZ3 protein (the third PDZ domain of PSD-95) that is known to display dynamic allostery, i.e., allostery without major structural changes. PDZ domains are one of the most abundant and evolutionarily conserved protein-protein interaction modules that take part in numerous cellular and biological functions, including dimerization and recognition of specific sequences of C terminus tails of other proteins [12]. The conserved structures in PDZ domains are six beta strands: $\beta1-\beta6$, and two alpha helices: $\alpha1$ and $\alpha2$. Some of the PDZ domain proteins have additional structures, which have been proposed to affect ligand binding. PDZ3 has $\alpha3$, $\beta7$, and $\beta8$ extensions [116]. It was shown that the removal of the $\alpha3$ domain results in a 21 times decrease in the binding affinity of PDZ3 [117], without a significant conformational change. This points to the allosteric nature of the PDZ3 domains, as well as to the entropic (as opposed to structural) nature of this allosteric behavior. The aligned structures of the 110-residue-long apo and holo forms of PDZ3 (PDB ID's 1bfe and 1be9, respectively) are shown in Figure 5.5. The ligand, which is represented in ball-and-stick form, is a 5-residue-long C-terminal segment of the Cysteine-rich PDZ-binding protein CRIPT. The binding pocket lies between the $\alpha2-\beta2$ regions, and is comprised of residues $[320-328]$ and $[371-380]$. Figure 5.6 presents the sequence and secondary structure assignment information.

First, the analysis techniques proposed in this chapter are validated in Section 5.4.1 by examining the displacements due to noise forces alone. This verification is based on the equivalence of the B factor values obtained via the proposed work flow and previously available computational techniques. Since the proposed techniques emphasize frequency domain representations, the frequency distribution of equilibrium fluctuations is obtained as an intermediate step. Then, in Section 5.4.2,

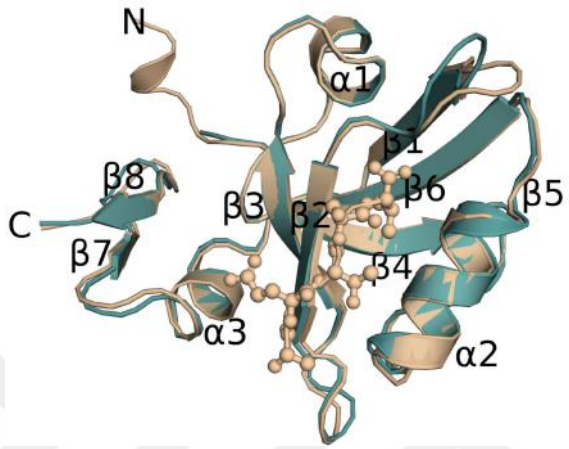


Figure 5.5: The aligned structures of the apo and holo forms of the PDZ3 protein. Wheat (teal) color is for the holo (apo) form. The ligand is represented in ball-and-stick format. PyMOL was used for the visualization [2].

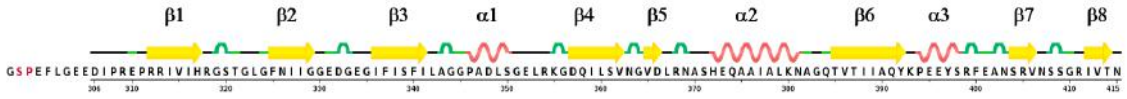


Figure 5.6: The secondary structure assignments from DSSP (definition of secondary structure of proteins) [3] of the holo form of the PDZ3 protein (PDBID: 1be9). Image is from the RCSB PDB (www.rcsb.org) [4].

we present frequency-decomposed SNR and channel capacity based analyses in the presence of external perturbations using both the *per residue scan* and *BP excitation* scenarios.

5.4.1 Without external force: equilibrium fluctuations

Mean square fluctuations (*MSFs*), which are proportional to experimentally measurable B factors, can be directly calculated from the pseudo-inverse of the Hessian of the potential energy function (please refer to Equations 5.45 and 5.47). We proposed an alternative frequency domain method to compute these equilibrium fluctuations by integrating the power spectral densities over the whole frequency range. The proposed scheme is based on the (linearized) Langevin formalism (please see Equation 5.1), which uses $\boldsymbol{H}$. Hence, the *MSFs* obtained from the pseudo-inverse of $\boldsymbol{H}$ and the *MSFs* from the proposed method should correspond to each other. Here, we consider $\boldsymbol{H}$ obtained both from MD ($\boldsymbol{H}^{\mathrm{MD}}$) and also from ANM ($\boldsymbol{H}^{\mathrm{ANM}}$) (see Section 3 for details). Figure 5.7 presents the B factors obtained for the apo form (PDBID:1bfe): from (i) the direct pseudo-inverse approach, (ii) the proposed PSD-based scheme, and (iii) experimentally measured values. The orange and green lines present B factors obtained from $\boldsymbol{H}^{\mathrm{MD}}$ and $\boldsymbol{H}^{\mathrm{ANM}}$, respectively. The blue and red lines represent B factors calculated from the integral of the PSDs both with $\boldsymbol{H}^{\mathrm{MD}}$ and $\boldsymbol{H}^{\mathrm{ANM}}$, respectively (please refer Section 5.2.6 for the details of the method). The perfect match between the proposed frequency domain scheme and the pseudo-inverse-$\boldsymbol{H}$ technique confirms the validity of the proposed work flow that utilizes the transfer functions with constraints introduced to overcome the singularity problem, and PSD integrations. Moreover, the agreement of the B factors between ANM and MD (except for the residues $[379-381]$) based Hessians justifies the use of simplified models for fluctuation analyses around an equilibrium structure. The black line in Figure 5.7 corresponds to the experimental B factors, which are reported in the associated PDB file. The secondary structure information is displayed at the bottom. The frequency decomposition of the *MSFs* are shown in Figure 5.8 for $\boldsymbol{H}^{\mathrm{MD}}$ and $\boldsymbol{H}^{\mathrm{ANM}}$. (Please see Figure 8.18 for the side views). In both cases, the low frequencies

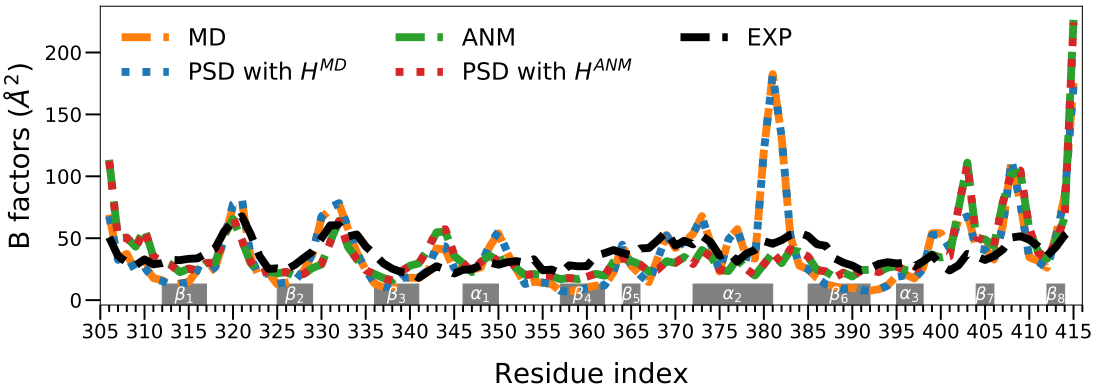


Figure 5.7: B factor values that are calculated with various methods for the apo form (PDBID:1bfe). The blue (red) lines show the values calculated via power spectral density integrations using $\boldsymbol{H}^{\mathrm{MD}}$ ($\boldsymbol{H}^{\mathrm{ANM}}$). The orange (green) line is calculated directly from the pseudo-inverse of $\boldsymbol{H}^{\mathrm{MD}}$ ($\boldsymbol{H}^{\mathrm{ANM}}$). The black line shows the experimental values. The values are scaled to correspond to the experimental ones.

contribute significantly more to the total *MSFs*. Also, there is an apparent trend: PSDs decrease as frequency increases, yet with varying rates for different residues. The PSDs obtained from $\boldsymbol{H}^{\mathrm{MD}}$ display a more rapid decrease starting from low frequencies, whereas for $\boldsymbol{H}^{\mathrm{ANM}}$, this decrease starts at higher frequency values. Although the difference in total *MSFs* for the two Hessians is relatively small, as seen in Figure 5.7, the spectral decompositions exhibit substantial differences. Hence, the equilibrium analysis suggests that the results from simplified models should be interpreted with caution.

5.4.2 External force excitations

5.4.2.1 Per residue scan

Figure 5.9 presents the capacity values obtained with *per residue scan* for $\boldsymbol{H}^{\mathrm{MD}}$. The normalized channel capacity values presented in panel (a) suggests a coupling between the first several residues closest to the C terminal and (β_7 & β_8) domains, as well as between β_1 and β_6, β_2 and (β_3 & β_4), β_5 and β_6, and β_7 and β_8 domain

(a) $\boldsymbol{H}^{\mathrm{MD}}$ (b) $\boldsymbol{H}^{\mathrm{ANM}}$

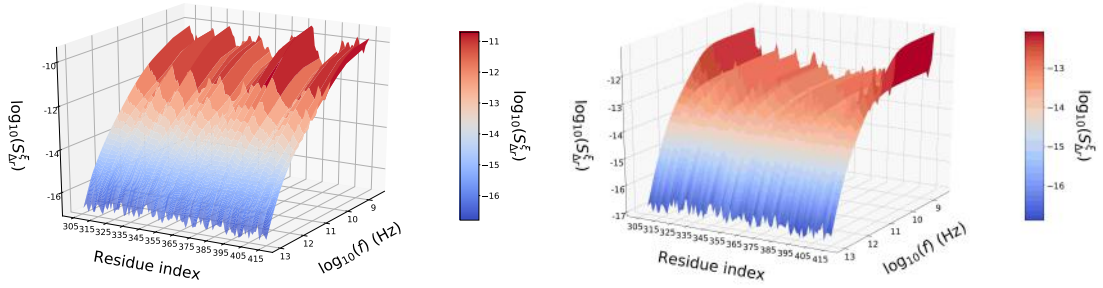


Figure 5.8: PSD profile of the equilibrium fluctuation with (a) $\boldsymbol{H}^{\mathrm{MD}}$ and (b) $\boldsymbol{H}^{\mathrm{ANM}}$.

pairs. Due to the fact that high channel capacity is expected between residue pairs that are close to each other in space, the capacity values are distance-weighted to detect the long-range interactions more clearly, as shown in Figure 5.9(b). The pairwise distance values that are used in this weighting procedure are calculated at the equilibrium point (please see Figure 8.16 for the pairwise distances). Despite being distant, the residue pairs that are found to be linked are listed in Table 5.1. The motions of SER 320 - SER 408, SER 409, and ASN 415 appear to be coupled. The results with $\boldsymbol{H}^{\mathrm{ANM}}$ for *per residue scan* (as well as for *BP excitation*) are presented in Figure 8.17.

5.4.2.2 Binding pocket (BP) excitation

Figure 5.10 presents the channel capacity values obtained with *BP excitation* for $\boldsymbol{H}^{\mathrm{MD}}$. Panel (a) presents the distance-weighted and normalized channel capacity values, while panel (b) presents a complementary visualization of the 3D protein structure without applying the weighting (The capacity values utilized in panel (b) are shown in Figure 8.19 in the same format of panel (a)). Please note that the residues that are located in the BP, which are represented by gray boxes in the left panel, are excluded from the normalization procedure since the external forces are

Input	-	Output
ASP 306	-	SER 320, THR 321
SER 320	-	ASP 332, ASN 403, THR 414, ASN 415
ASN 407	-	SER 320, THR 321
SER 408	-	SER 320, THR 321, ASN 381
SER 409	-	SER 320
ASN 415	-	SER 320, THR 321

Table 5.1: Residue pairs that are located far apart in space, yet have high channel capacity values that are identified by *per residue scan*, with $\boldsymbol{H}^{\mathrm{MD}}$.

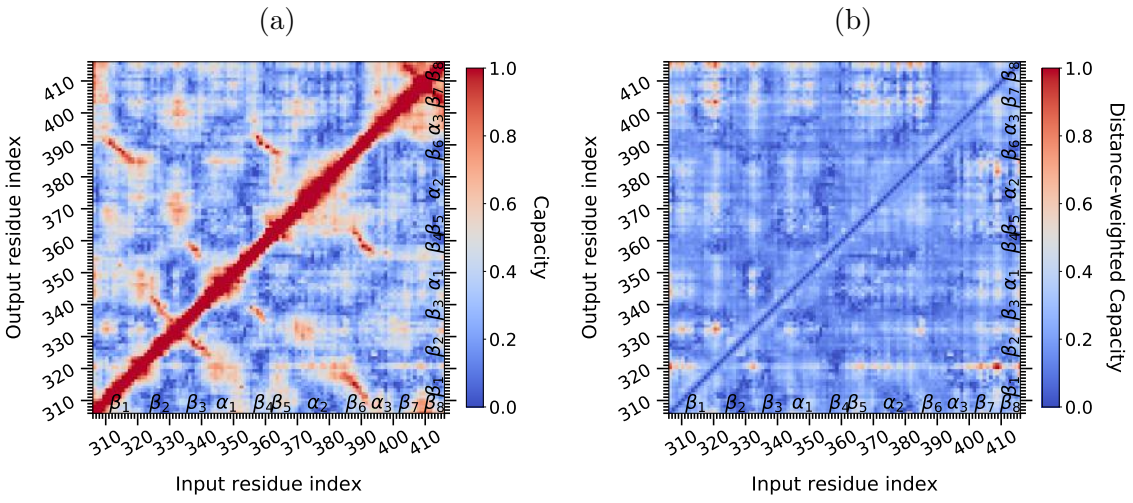


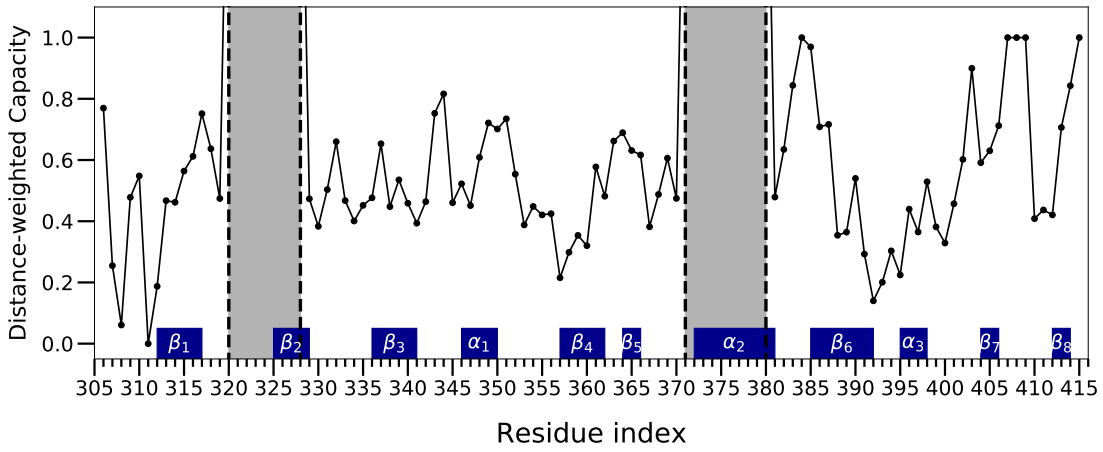
Figure 5.9: Per residue scan with $\boldsymbol{H}^{\mathrm{MD}}$. (a) Normalized, (b) Distance-weighted and normalized channel capacity values.

directly applied to them. According to the top panel, ASP 306, HIS 317, ALA 343, GLY 344, GLY 383, GLN 384, THR 385, ASN 403, ASN 407, SER 408, SER 409, THR 414, and ASN 415 are identified to have the potential to display the most significant allosteric response to the ligand-binding event.

Figure 5.11 presents the residues that have been reported as *critical* using various computational and experimental techniques, which are also listed in Table 8.3. The normalized channel capacity values (without distance-weighting) obtained from *BP excitation* are displayed as color coded at the bottom row. As in the previous studies, the BP residues are excluded from the analysis. As can be seen from the figure, there is little agreement among the methods that aim to identify the critical residues involved in the PDZ3 protein. Nonetheless, there are several residues that most of the methods agree on, such as GLY 329, ILE 338, ILE 341, ALA 347, LEU 353, VAL 362, and VAL 386. These residues (with the exception of LEU 353) do have a high capacity value based on the results obtained with our channel capacity analysis method. In addition to those listed here, the rest of the residues that have high capacity values were also reported to be critical by at least several previous techniques. Even though these findings point to the success of the proposed method, it is not possible to assess the degree of accuracy with respect to a golden reference due to the lack of consistency among previous methods. It is unfortunate that there is no consensus in the literature on the residues that play a significant role in the allosteric behavior of this protein. However, reaching a consensus seems to be exceptionally difficult when the tremendous challenges involved in both experimental and computational techniques aimed at deciphering allostery are considered. Nevertheless, we believe that the proposed channel capacity technique provides a fresh approach for deciphering the mechanisms of allostery, and could help detect hidden allosteric interactions.

In order to investigate role of frequency in determining the allosteric responses, we present the frequency dependent SNR profiles in Figure 5.12(a). The side views of the figure are in Figure 8.20(a). The frequency dependence of the SNRs for the individual residues are also shown in Figure 8.21. Although the majority of the

(a)



(b)

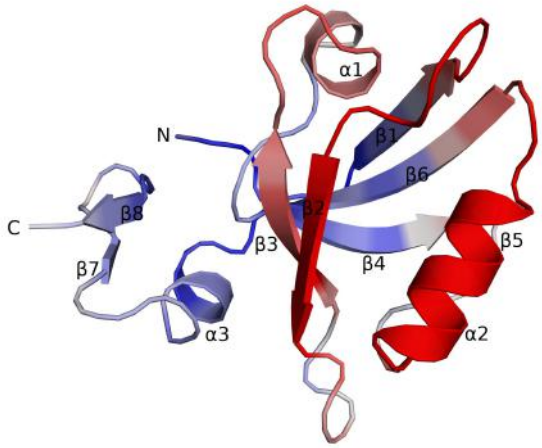


Figure 5.10: BP excitation with $\boldsymbol{H}^{\mathrm{MD}}$ (b) Distance-weighted and normalized channel capacity values. The gray area corresponds to the binding pocket residues upon which the force is applied. (b) Apo protein structure (PDBID:1bfe) is colored according to the capacity values (without distance weighting): red indicates the highest capacities and blue is for the lowest.

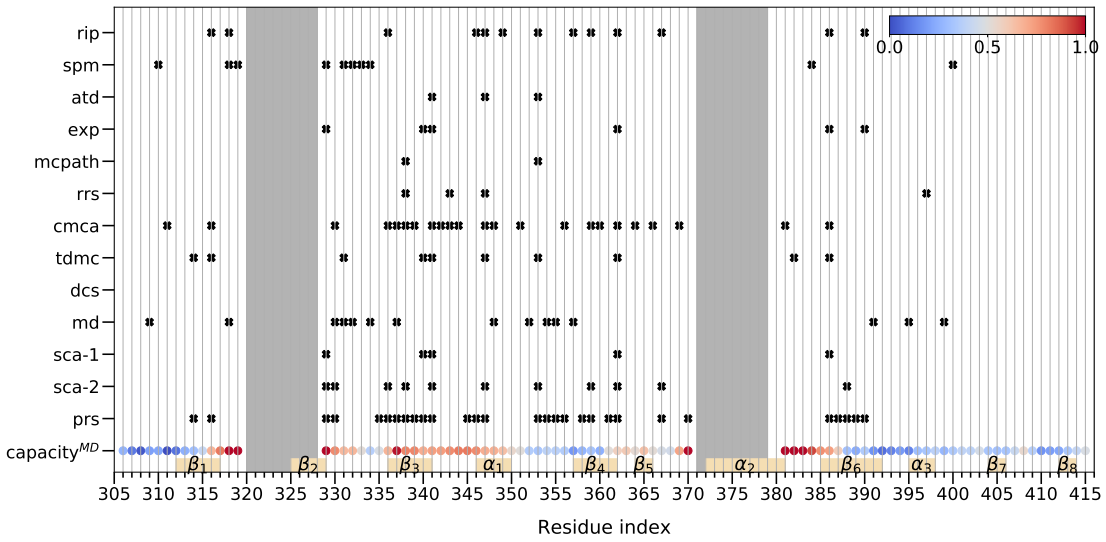


Figure 5.11: Capacity values from *BP excitation* and the critical residues identified by various methods. The gray areas indicate the residues on which the external force is applied in the capacity analysis. Capacity$^{\mathrm{MD}}$ refers to the results obtained with $\boldsymbol{H}^{\mathrm{MD}}$. Abbreviations: prs - perturbation response scanning [5], exp - experimental [6], sca-1 - statistical coupling analysis [7], sca-2 - statistical coupling analysis [8], atd - anisotropic thermal diffusion [9], spm - structural perturbation method [10], rip - rotamerically induced perturbation [11], md - molecular dynamics [12], dcs - deep coupling scan [13], tdmc - thermodynamic double mutant cycle [14], cmca - conservation mutation correlation analysis [15], rrs - rigid-residue scan [16], mcpath - Monte Carlo path [17].

residues display monotonic decrease in SNR with increasing frequency, it is worth noting that some of the residues exhibit a peak, a sort of *resonance*, around a certain frequency value, that is specific to the residue. The residues with distinct frequency response are presented in Figure 5.12(b). The responses of the residues differ in terms of the magnitude of SNR, the slope of the decay with frequency, and the existence and frequency location of a resonance at higher frequencies. For instance, GLY 329, which is one of the few consistently identified critical allosteric residues in Figure 5.11 has a relatively high response until the high end of the spectra. The resonance around 2 THz is also worth noting for GLY 329. The resonance at about the same frequency is significantly more pronounced for GLN 384. The other listed residues display resonances with smaller magnitudes at high frequencies. With these results, we underline that the frequency decomposed SNR analysis provides a more detailed picture of allostery, paving the way for gaining a better understanding of allosteric mechanisms.

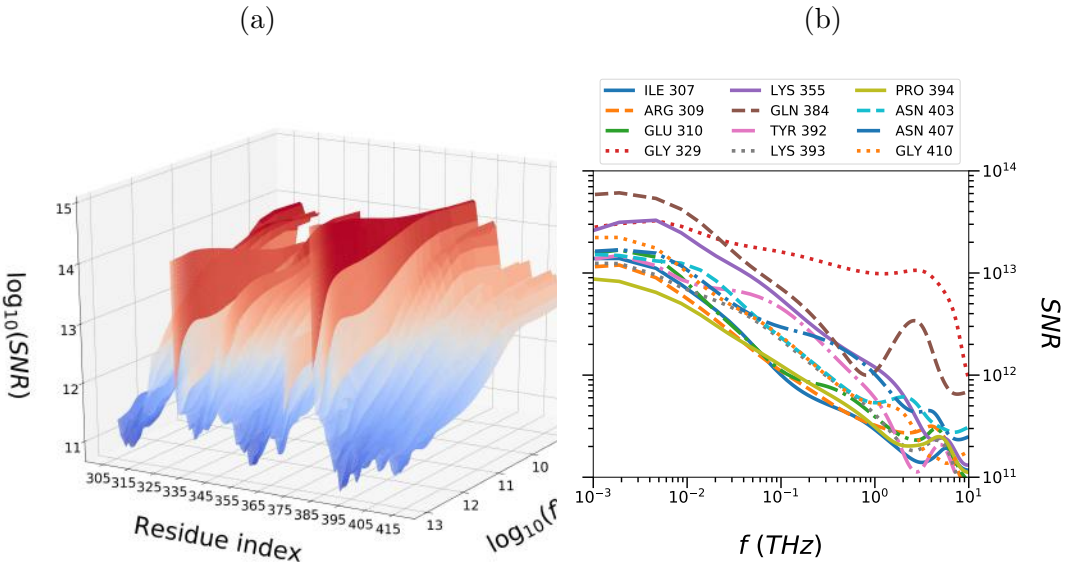


Figure 5.12: (a) Signal-to-Noise ratio (SNR) with *BP excitation* using $\boldsymbol{H}^{\mathrm{MD}}$, (b) Selected residues with characteristic frequency response.

Part II

Representation Learning

Chapter 6

INTERPRETABLE EMBEDDINGS FROM MOLECULAR SIMULATIONS USING GAUSSIAN MIXTURE VARIATIONAL AUTOENCODERS

6.1 Introduction

Particle-based computer simulations can provide unprecedented mechanistic insight into the driving forces of complex molecular systems, in contexts ranging from biochemistry to materials science [119, 18, 120]. These simulations rely on numerical integration of the relevant equations of motion as a means to navigate the system's conformational space. Due to the high dimensionality of this space, which prevents the exhaustive enumeration of all microstates, exploration is typically achieved through importance sampling [108]. Conformational sampling leads to an estimate of the potential energy landscape (PEL), which follows a Boltzmann distribution at equilibrium. Unfortunately, characterization of the PEL suffers from the so-called *curse of dimensionality* [121]—organization of the data in the high-dimensional space is challenging due to low population density. This problem is often remedied by projecting the PEL onto a lower-dimensional manifold, i.e., by performing a dimensionality reduction. By averaging over presumably unimportant degrees of freedom, the resulting low-dimensional surface represents a *free*-energy landscape (FEL). The ideal FEL distinguishes between microstates that are separated by large barriers on the PEL, yielding a partitioning of configuration space into collections of microstates, i.e., metastable basins. If all the largest barriers are accounted for, intra-basin diffusion will occur much faster than inter-barrier crossing events, allowing an accurate, albeit coarse-grained, description of both the static

The content of this chapter has previously appeared in [118].

and dynamical properties of the system. There is a long history of methods for finding an optimal low-dimensional representation from a given set of data, employing both linear (e.g., principal component analysis [122], time-lagged independent component analysis [123]) and nonlinear (e.g., Isomap [124], diffusion map [125], and Sketchmap [126]) transformations.

In the last couple of years, there has been a growing interest in applying (deep) neural networks to automate the discovery of CVs [38, 39, 40, 41, 42]. One architecture that stands out as conceptually appealing is the autoencoder [127]. An autoencoder is a bow-tie-shaped network that forces an information compression in the bottleneck region. While the first half of the network (the encoder) reduces the input to a predefined lower dimension, the second half (the decoder) aims at transforming from the low-dimensional to the original representation. The weights of the neural network are tuned to minimize an objective or *loss* function, which typically penalizes deviations between input and output data. As such, the autoencoder aims at discovering a *latent space* (embedding) that faithfully describes the essential features of the high-dimensional input data. This makes autoencoders well suited for constructing low-dimensional FELs from molecular simulation data [43, 38, 44].

Traditional autoencoders lack continuity in the latent space, preventing interpolation between training points and, thus, its generative ability. Variational autoencoders (VAEs) remedy this limitation by modeling the input probability distribution using Bayesian inference [128]. VAEs enable sampling new data from the learned distribution (i.e., VAEs are generative models), and are also well suited to provide interpretable and disentangled data representations in the low-dimensional space [129]. Within the VAE framework, the latent distribution is forced to resemble a predefined probability distribution, called the *prior*. Although the VAE framework does not impose any particular prior distribution, it is often chosen as a normal distribution for computational convenience. This prior induces an "anti-clustering" effect in the latent space, which can prohibit the identification of meaningful clusters and impede the construction of optimal FELs from molecular simulations. The autoencoder-based approaches were recently extended to explicitly incorporate the

temporal nature of the data via a time lag in the network architecture [130, 131]. These time-lagged autoencoders aim to retain information about the slowest dynamical modes sampled in the underlying simulation trajectory and, as a consequence, may encourage metastable clustering in the latent space. However, they are also limited in terms of characterizing the hierarchy of long timescale processes [132], and only indirectly address the anti-clustering issue.

In this work, we propose to directly acknowledge the multi-basin structure of an ideal FEL by employing a Gaussian mixture model [133] as the prior distribution for the VAE latent space. The resulting *Gaussian mixture variational autoencoder* (GMVAE) retains the computational ease and reconstruction fidelity of traditional VAEs, while enforcing a more faithful description of the underlying physics: the resulting FEL clearly distinguishes between metastable basins separated by large free-energy barriers. We demonstrate the benefits of the GMVAE approach through explicit comparisons with the traditional VAE for two widely-studied toy models and for the standard benchmark system for conformational dynamics, alanine dipeptide, as well as a more challenging disordered peptide ensemble. To ensure the presence of distinct distributions in the latent space, the GMVAE introduces a categorical variable that (probabilistically) assigns each input configuration to the set of clusters. Thus, the GMVAE simultaneously performs dimensionality reduction and unsupervised clustering. Remarkably, the GMVAE clustering is capable of identifying the inherent dimensionality of the input data, in terms of the number of Gaussians required to categorize the data. In the case of hierarchical input data (i.e., data with distinct dimensionality depending on the level of resolution), we show that the GMVAE makes a reasonable prediction for the number of clusters, independent of the given hyperparameter, based on the dimensionality of the latent space and characteristics of the data. Beyond the representation of static equilibrium properties, by constructing MSMs from the GMVAE embedding, we show that our approach is also a promising avenue for accurately describing the long timescale *dynamical* properties of the data. In contrast to recent deep neural network approaches that aim to directly model the *propagator* of the system's dynamics [134, 135], the con-

struction of MSMs from the learned FEL offers a different strategy: explicitly testing to what extent a representation appropriate for the statics is directly amenable for the dynamics.

6.2 Theory and methods

6.2.1 Autoencoder

Autoencoders are special types of neural networks that are used for the task of representation learning in an unsupervised manner. They are composed of two connected parts: the encoder compresses the input signal to a low-dimensional representation, whereas, the decoder aims to reconstruct the input at full dimensionality from the reduced-space representation. The *reconstruction loss*, usually defined as either the mean-squared error or cross-entropy between the input, x, and the output, x', is minimized via backpropagation. Since the bottleneck dimension is typically much less than the original dimension, autoencoders learn the most compact representation of the input. Furthermore, because neural networks are universal function approximators, the learned data projections can generally preserve much more of the relevant information than with PCA or other basic linear projection techniques. Figure 6.1 shows the schematic structure of an autoencoder with mean-squared error loss. There are different types of autoencoders which are tailored for special tasks. For instance, sparse autoencoders impose sparsity constraints during optimization, whereas convolutional autoencoders utilize convolutional layers instead of fully-connected layers, in which case they learn the optimal filters. Variational autoencoders, which model the latent space probabilistically, are used for generative purposes, i.e., they can create new samples that look like the ones in the training dataset without simple data replication.

6.2.2 Variational autoencoder (VAE)

Variational autoencoders were introduced in [128]. In general, the theory of VAEs is approached from two different perspectives: variational inference and neural net-

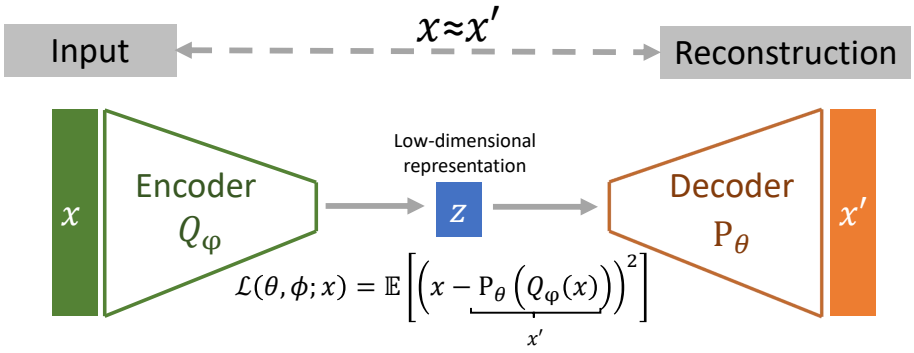


Figure 6.1: Schematic of an autoencoder architecture with mean-squared error reconstruction loss.

works. This section starts with the former interpretation and then illustrates the connection between them. We mostly follow the notation and reasoning used in [136]. The input data and the latent variable are denoted by x and z, respectively.

The objective of the VAE is to find the posterior distribution $P(z|x)$, which can be written in terms of the likelihood $P(x|z)$, the prior $P(z)$, and the marginal probability density of x, $P(x)$, using Bayes law as

$$P(z|x) = \frac{P(x|z)P(z)}{P(x)} \ . \tag{6.1}$$

The denominator $P(x)$ is called the evidence and it could, in principle, be calculated using

$$P(x) = \int \mathrm{d}z \, P(x|z)P(z) \ , \tag{6.2}$$

once the prior is selected. However, the calculation is typically intractable, as it needs to be evaluated over all configurations of the latent variable z. Therefore, the posterior is approximated using *variational inference* with a chosen easy-to-evaluate family of distributions $Q_\phi(z|x)$, e.g., Gaussian functions, where ϕ is the variational parameter of the distribution. In particular, $P(z|x)$ is inferred using $Q_\phi(z|x)$ by reformulating the problem within an optimization framework, such that the Kullback-Leibler divergence between $Q_\phi(z|x)$ and $P(z|x)$ is minimized. The KL

divergence between Q and P is defined as

$$\begin{aligned} D_{\mathrm{KL}}[Q_\phi(z|x)||P(z|x)] &= \sum_z Q_\phi(z|x)\, \log \frac{Q_\phi(z|x)}{P(z|x)} \\ &= \mathbb{E}\left[\log \frac{Q_\phi(z|x)}{P(z|x)}\right] \\ &= \mathbb{E}[\log Q_\phi(z|x) - \log P(z|x)] \ . \end{aligned} \tag{6.3}$$

Equation 6.1 is then inserted into the posterior definition

$$\begin{aligned} D_{\mathrm{KL}}[Q_\phi(z|x)||P(z|x)] &= \mathbb{E}\left[\log Q_\phi(z|x) - \log \frac{P(x|z)P(z)}{P(x)}\right] \\ &= \mathbb{E}[\log Q_\phi(z|x) - \log P(x|z) - \log P(z) + \log P(x)] \ . \end{aligned} \tag{6.4}$$

Since the expectation is taken over z, $P(x)$ can be moved out of the expectation

$$D_{\mathrm{KL}}[Q_\phi(z|x)||P(z|x)] - \log P(x) = -\underbrace{\mathbb{E}[\log P(x,z) - \log Q_\phi(z|x)]}_{\mathrm{ELBO}(\phi)} \ . \tag{6.5}$$

The initial objective of minimizing the KL divergence between the exact and the approximate posterior is equivalent to maximizing the ELBO (Evidence Lower BOund), defined in Equation 6.5.

Equation 6.5 can also be rewritten in terms of a different KL divergence:

$$D_{\mathrm{KL}}[Q_\phi(z|x)||P(z|x)] - \log P(x) = D_{\mathrm{KL}}[Q_\phi(z|x)||P(z)] - \mathbb{E}[\log P(x|z)] \ . \tag{6.6}$$

Here the neural network perspective comes into play, as depicted schematically in Figure 6.2(a). $Q_\phi(z|x)$ acts like an encoder (*inference*), and transforms the data into the latent variable z. On the other hand, $P(z|x)$ (which can also be parametrized with the network parameter θ as $P_\theta(z|x)$[1]) generates the data from the latent representation, analogous to a decoder (*generator*). The parameters correspond to the weights and biases of the neural networks. Note that the initial aim is to minimize $D_{\mathrm{KL}}[Q_\phi(z|x)||P(z|x)]$, which is equivalent to minimizing the RHS of Equation 6.6. The first term enforces the encoder to be similar to the chosen prior $P(z)$, which acts as a regularization, whereas the second term on the RHS deals with how well the reconstructions match the original input.

[1]Both of the notations are used interchangeably.

6.2.2.1 Standard selections for the family of inference distributions and for the prior distribution

In order to use Equation 6.6 in an optimization procedure, both the family of distributions for inference, $Q_\phi(z|x)$, as well as the prior distribution, $P(z)$, must be specified. The most common assumption is that $Q_\phi(z|x)$ ($P(z)$) is a unimodal Gaussian distribution with mean $\mu(x)$ ($\mathbf{0}$) and diagonal covariance $\Sigma(x)$ ($\mathbf{1}$). Then, $D_{\mathrm{KL}}[Q_\phi(z|x)||P(z)]$ has a closed form solution:

$$\begin{aligned} D_{\mathrm{KL}}[Q_\phi(z|x)||P(z)] &= D_{\mathrm{KL}}[\mathcal{N}(\mu(x),\Sigma(x))||\mathcal{N}(0,\mathbf{1})] \\ &= \frac{1}{2}\left(\mathrm{tr}(\Sigma(x)) + \mu(x)^T\mu(x) - d - \log\,\det(\Sigma(x))\right) \ , \end{aligned} \tag{6.7}$$

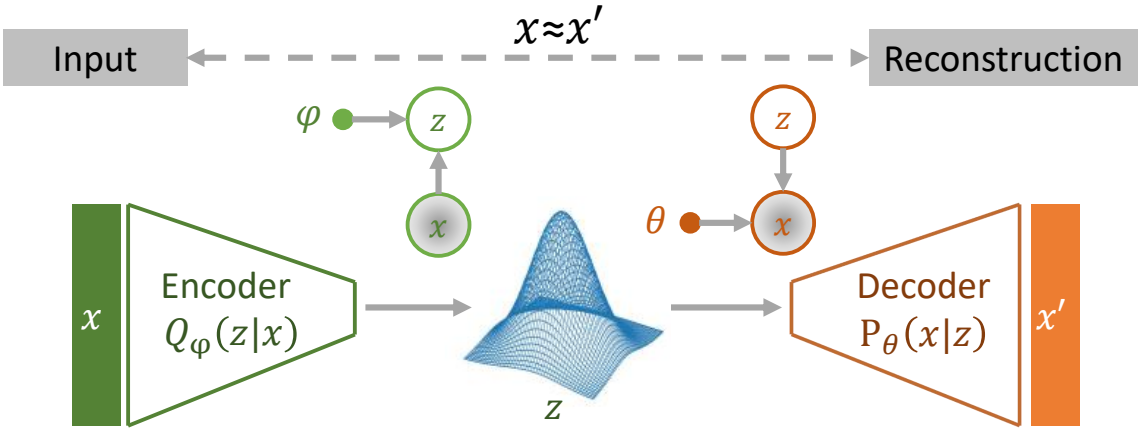
where d is the dimension of the Gaussian and tr denotes the trace. Although the unimodal Gaussian assumption simplifies the calculations, it also restricts the possible latent space representations, and may hinder the performance of the variational autoencoder by pushing the latent space to be described by highly-overlapping clusters.

6.2.3 Gaussian mixture variational autoencoder

This section is largely distilled from the discussion and insights presented in [137]. The term Gaussian mixture variational autoencoder is open to misinterpretations. There exist several distinct architectures given this name, with variations in the choice of generative or inference models [133, 138, 139, 140]. In the present work, we take both the approximate posterior, (i.e., the family of distribution functions for inference), $Q_\phi(y,z|x)$, and the latent space distribution (i.e., the prior), $P(z)$, to be Gaussian mixtures. Note that we have introduced a categorical variable, y, which identifies which Gaussian each particular data point belongs to. The inference model can be written as

$$Q_\phi(y,z|x) = Q_\phi(y|x)Q_\phi(z|x,y) \ . \tag{6.8}$$

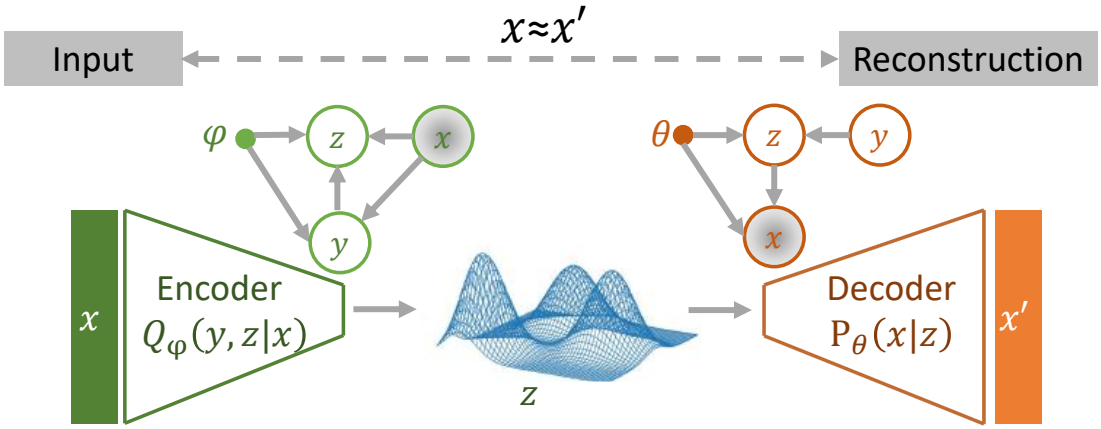
The latent space is composed of k distinct Gaussians, i.e., $Q_\phi(z|x,y_i)$ is assumed to be Gaussian, where $i \in 0,1,\ldots,k-1$. Thus, the approximate posterior becomes a Gaussian mixture.



$$\mathcal{L}(\theta,\phi;x) = \mathbb{E}_{Q_\phi(x,z)}\left[\log P_\theta(x,z) - \log Q_\phi(z|x)\right]$$

$$\mathcal{L}(\theta,\phi;x) = \underbrace{\mathbb{E}_{Q_\phi(z|x)}[\log P_\theta(x|z)]}_{\text{Reconstruction}} - \underbrace{\mathrm{D}_{\mathrm{KL}}\left(Q_\phi(z|x)||P(z)\right)}_{\text{Regularization}}$$

(a) Schematic of a variational autoencoder with unimodal Gaussian prior.



$$\mathcal{L}(\theta,\phi;x) = \mathbb{E}_{Q_\phi(y,z|x)}\left[\log P_\theta(x,y,z) - \log Q_\phi(y,z|x)\right]$$

$$\mathcal{L}(\theta,\phi;x) = \mathbb{E}_{Q_\phi(y,z|x)}\Big[\underbrace{\log\frac{P(y)}{Q_\phi(y|x)}}_{\text{Entropy}} + \underbrace{\log\frac{P_\theta(z|y)}{Q_\phi(z|x,y)}}_{\text{Regularization}} + \underbrace{\log P_\theta(x|z)}_{\text{Reconstruction}}\Big]$$

(b) Schematic of a Gaussian mixture variational autoencoder.

Figure 6.2: (a) The VAE and (b) GMVAE architectures. In the probabilistic graph representation, circle nodes represent the random variables, and directed edges represent statistical dependencies between the variables in the two ends. Dot nodes are used to indicate the parameters of the model, while some of the nodes are intentionally filled to differentiate the observed random variables from the non-observed ones which are left empty.

Similar to Equation 6.5, the ELBO can be written as

$$\mathrm{ELBO_m} = \mathbb{E}_{Q_\phi(y,z|x)}[\log P_\theta(x,y,z) - \log Q_\phi(y,z|x)] \ , \tag{6.9}$$

where the number of Gaussians, k, is a hyperparameter, and the subscript m is used to distinguish $\mathrm{ELBO_m}$ from the VAE ELBO. $P_\theta(x,y,z)$ can be written as $P_\theta(x,y,z) = P_\theta(x|y,z)P_\theta(z|y)P(y)$ using conditioning without any assumptions. Then, by assuming that x is conditionally independent of y, i.e., $P_\theta(x|y,z) = P_\theta(x|z)$ (see the graph representation in Figure 6.2(b)), the joint probability can be expressed as

$$P_\theta(x,y,z) = P_\theta(x|z)P_\theta(z|y)P(y) \ . \tag{6.10}$$

By inserting Equations 6.8 and 6.10 into Equation 6.9, ELBO_m becomes

$$\begin{aligned}\mathrm{ELBO_m} &= \mathbb{E}_{Q(y,z|x)}[\log P(y)P_\theta(z|y)P_\theta(x|z) - \log Q_\phi(y|x)Q_\phi(z|x,y)] \\ &= \mathbb{E}_{Q(y,z|x)}\Big[\log P(y) - \log Q_\phi(y|x) + \log \frac{P_\theta(z|y)}{Q_\phi(z|x,y)} + \log P_\theta(x|z)\Big] \ .\end{aligned} \tag{6.11}$$

Similar to the VAE, the third and fourth terms represent regularization and reconstruction contributions to the loss, respectively. The initial prior on y is selected as a uniform multinomial distribution, while $\mathbb{E}_{Q(y,z|x)}[\log Q_\phi(y|x)]$ can be interpreted as a conditional entropy, reflecting how informative x is on y. To directly control the impact of the clustering relative to the other loss terms during training, we introduced a weighting factor, α, on the mutual information between x and y:

$$\mathrm{ELBO_m} = \mathbb{E}_{Q(y,z|x)}\Big[\log P(y) - \alpha \log Q_\phi(y|x) + \log \frac{P_\theta(z|y)}{Q_\phi(z|x,y)} + \log P_\theta(x|z)\Big] \ . \tag{6.12}$$

Figure 6.3 presents a more detailed schematic of the GMVAE architecture, while Table 6.1 presents a summary of the probability distributions utilized in the model. First, data points are probabilistically assigned to k clusters ($\mathrm{NN(Q}_y\mathrm{)}$). $Q(y|x)$ represents these cluster assignment probabilities, and has a multinomial distribution. Since each cluster is assumed to have Gaussian distribution in the latent space,

the mean and variance of each of these Gaussians ($Q(z|x,y)$) are learned via the encoder part of the neural network (NN(Q_z)). The low-dimensional representation, z, is then obtained by first sampling and then taking the expected value of these samples, i.e., $z = \sum_{i=0}^{k-1} p(y_i|x)z_i$. As the first step in decoding, the moments of the corresponding low-dimensional representation z is learned by NN(P_z) from each Gaussian-distributed individual cluster y_i, which is then followed by a sampling operation. $P(y)$ in the decoder is assumed to be uniformly distributed among the k clusters. Next, using the encodings, z_i's, the associated x reconstructions are obtained again by sampling from the x' by the NN(P_x). Similar to the encoder, the decoder obtains a fixed reconstruction by taking the expected value of x'_i's.

$$
\begin{aligned}
Q(z|x,y) &= \mathcal{N}(\mu_z(x,y), \sigma_z^2(x,y)) \\
Q(y|x) &= \text{Multinomial}(f(x))
\end{aligned}
\qquad
\begin{aligned}
P(y) &= \text{Uniform}(\tfrac{1}{k}) \\
P(z|y) &= \mathcal{N}(\mu_z(y), \sigma_z^2(y)) \\
P(x|z) &= \mathcal{N}(\mu_x(z), \sigma_x^2(z))
\end{aligned}
$$

Table 6.1: Distributions in the GMVAE model. Left (right) column corresponds to the distributions in the encoder (decoder) part.

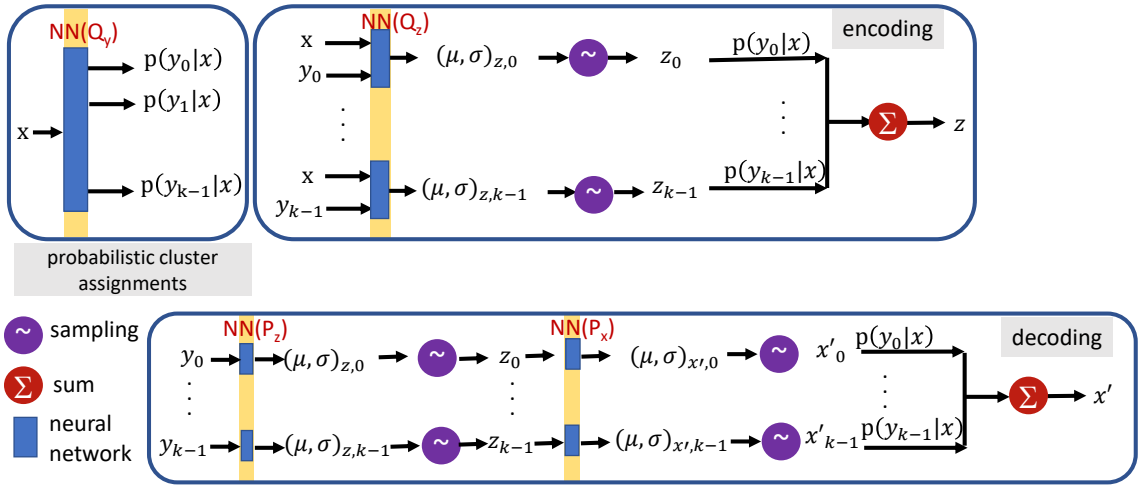


Figure 6.3: Schematic of the GMVAE workflow.

6.2.3.1 Determination of cluster labels and thresholding scheme

The clustering within the GMVAE is probabilistic, i.e., each data point is assigned membership probabilities (between 0 and 1) to each of the clusters. Since most configurations are assigned predominantly to a single cluster, we perform a hard cluster assignment by assigning each data point to the cluster with highest membership probability. However, in cases where a configuration has similar membership probabilities for multiple clusters, this simple assignment may introduce errors when determining properties (e.g., transition probabilities) of the clusters. Thus, we also considered a different approach by enforcing a thresholding value for cluster assignment. More specifically, each configuration is only assigned to a cluster if the largest membership probability is above a chosen cut-off value. A naive coring scheme followed the thresholding operation such that the points that had been identified as noise were assigned back to their previous cluster index for all other dynamical analyses.

6.2.3.2 GMVAE architecture and training hyperparameters

The GMVAE algorithm was implemented in Tensorflow [141]. Training was performed in all cases with fully-connected layers, using the Adam optimization algorithm [142]. The Softmax activation function was used for probabilistic cluster assignments, while ReLu activation functions were employed in all hidden layers. The means were obtained without any activation, whereas Softplus activation was employed to obtain the variances. Table 6.2 shows the values of the hyperparameters for each example system. Default values were employed wherever the parameters are not specified. The NN($\cdot$)'s correspond to the neural networks labeled in Figure 6.3. NN(Q_y) performs probabilistic cluster assignments, NN(Q_z) is for learning the moments of each Gaussian distribution in the encoding, whereas NN(P_z) and NN(P_x) are for the decoding of the z and x, respectively. The lengths of the "Number of nodes" entries correspond to the number of hidden layers. Hyperparameter optimization was carried out as follows. The number of nodes was initialized as [16, 16]. The number of nodes in the decoder (NN(P_x)) was then increased whenever a large

and non-decreasing reconstruction loss was observed. Our overall observation for the considered examples is that the learning rate and batch size should be kept relatively low to promote the formation of distinct clusters. The VAE results (with unimodal Gaussian prior) that are provided as comparison are obtained using $k = 1$, while keeping the remaining parameters equal to the values in the corresponding GMVAE model.

	1D 4-well	Müller-Brown	Dipeptide	$AAQAA_3$ - I	$AAQAA_3$ - II
# of clusters (k)	4	5	8	10	6
Input dimension (n)	1	2	25	60	126
Latent dimension (d)	1	1	2	2	2
# of nodes (NN(Q_y))	[16, 16]	[32]	[32]	[16, 16]	[128]
# of nodes (NN(Q_z))	[16, 16]	[16]	[16]	[16, 16]	[16]
# of nodes (NN(P_z))	[16, 16]	[16]	[16]	[16, 16]	[16]
# of nodes (NN(P_x))	[16, 16]	[128]	[128]	[16, 16]	[256]
α	0.5	0.05	0.05	0.3	0.95
Batch size	32000	5000	5000	10000	3000
Learning rate	0.00005	0.0001	0.00015	0.001	0.00005
# of epochs	50	400	100	300	2000
Probability cut-off	None	None	None	0.95	0.98

Table 6.2: Architecture specification and training hyperparameters

6.2.4 Markov state models

Markov state models (MSMs) represent the dynamics generated by a molecular simulation trajectory as a series of memoryless jumps between a discrete set of states [143]. Given a configuration-space discretization, a transition probability matrix, $\mathbf{P}(\tau)$, is obtained by counting the transitions between pairs of states within a given lag time, τ, and then performing a maximum likelihood optimization [144]. The eigenvalues of $\mathbf{P}(\tau)$, $\{\lambda_i(\tau)\}$, are related to characteristic timescales of the

system's dynamics:

$$t_i(\tau) = -\frac{\tau}{\ln|\lambda_i(\tau)|} \ , \tag{6.13}$$

where $t_i(\tau)$ is the timescale corresponding to the i^{th} eigenvalue, $\lambda_i(\tau)$. The time lag parameter τ is typically chosen by performing the "implied timescale test", which assesses the Markovianity of $\mathbf{P}(\tau)$ through the convergence of its timescales with increasing τ. In other words, $\{t_i(\tau)\}$ is plotted as a function of τ, and τ is then chosen as small as possible such that the largest timescales are *sufficiently* converged. Once τ is chosen, the accuracy of $\mathbf{P}(\tau)$ is determined via the Chapman-Kolmogorov (CK) test, which compares the estimated and predicted probability decay out of a given state. The predicted values are obtained using the CK equation, i.e., using the Markovian property of the model:

$$p_{ij}(m\tau) = p_{ij}^m(\tau) \ , \tag{6.14}$$

where $p_{ij}(\tau)$ is the probability of transitioning from state i to state j within time τ, and m is a positive integer. The CK test is often performed on metastables states of the system—collections of quickly interconverting microstates.

Within the standard Markov state modeling workflow, microstates are typically defined on low-dimensional projections of the full-dimensional configuration space. Therefore, obtaining a relevant transformation of the molecular simulation data is the key. To this end, time-lagged independent component analysis (TICA) [123, 145] is one of the most commonly used dimensionality reduction methods, as its objective is to maximize the autocorrelation of the data at the given lag time, making it especially well suited for kinetic modeling purposes. Metastable states are typically obtained via a dynamical coarse-graining procedure, e.g., PCCA$^+$ [146] whose objective is to retain an accurate description of the dominant eigenvectors of the transition probability matrix. The resulting metastable states are then used as representative collections of microstates for performing the CK test. In many cases, a coarse-grained MSM at the resolution of the metastable states is constructed, providing an easily interpretable, albeit often qualitative, picture of the long timescale processes.

In this study, the GMVAE performs the dimensionality reduction and clustering simultaneously, yielding a coarse-grained description of configuration space directly, without the need for further dynamical clustering. The (coarse-grained) MSMs are constructed from the discretized trajectories obtained using the simple cluster assignment based on the GMVAE membership probabilities as described in Section 6.2.3.1. MSM construction and analysis was performed using the PyEMMA package [147].

6.2.5 Peptide analysis

The helical propensity of the peptide was determined using the Lifson-Roig perspective, which assigns each residue to either a helical (h) or coil (c) state, according to the dihedral angles along the peptide backbone (i.e., the Ramachandran plot) [148, 149]. Therefore, the number of different conformations of the peptide is limited to 2^N, where N is the number of residues; $N = 15$ for $AAQAA_3$. The propensity of residue i to be part of a "helical segment", $\langle h_i \rangle$, is then defined as the probability that residue i as well as its two neighboring residues are simultaneously found in a helical state. The average fraction of helical segments, $\langle f_h \rangle$, is obtained by averaging $\langle h_i \rangle$ over all residue positions: $\sum_{i=0}^{N-1} \frac{1}{N} \langle h_i \rangle$. To distinguish between partial helical structures occuring at the N- and C-terminus ends of the peptide backbone, we define $\langle h_N \rangle = \sum_{i=1}^{6} \frac{1}{6} \langle h_i \rangle$ and $\langle h_C \rangle = \sum_{i=8}^{13} \frac{1}{6} \langle h_i \rangle$. Note that the terminus residue from each end is not taken into consideration.

The dRMSD measures the average deviation of internal distances from the corresponding distances in a reference structure, and is calculated as

$$\mathrm{dRMSD}(\mathbf{X}(t), \mathbf{X}^r) = \sqrt{\sum_{i \neq j} (||\mathbf{X}_i(t) - \mathbf{X}_j(t)|| - ||\mathbf{X}_i^r - \mathbf{X}_j^r||)^2} \,, \tag{6.15}$$

where $\mathbf{X}(t)$ represents the conformation at time t, $\mathbf{X}^r$ is the conformation for the reference structure, and $||\cdot||$ denotes the Euclidean norm. Note that, unlike other RMSD metrics, no pre-alignment of structures is required. In this study, due to the large fluctuations of the end residues, two residues from each end of the peptide were excluded in the dRMSD calculations. dRMSD was calculated using the positions of

the C_α atoms only. Helix, hairpin-like, and extended (coil) structures were separately considered as reference structures as illustrated in Figure 8.33.

6.3 Results

Variational autoencoders (VAEs) have been previously applied for dimensionality reduction of molecular simulation data [131, 39, 150]. VAEs typically employ a normal distribution to represent both the prior distribution in the latent space and the family of distributions for variational inference. In this work, we extend traditional VAEs by representing these distributions with Gaussian mixture models. The resulting Gaussian mixture VAE (GMVAE) adopts the physics-based viewpoint that an optimal embedding of the simulation data should give rise to a free-energy landscape (FEL) with well-separated clusters of configurations, which correspond to metastable states that are separated by large barriers along the high-dimensional potential energy landscape. The GMVAE introduces a categorical variable, y, which represents the various underlying Gaussian distributions to which each configuration will be (probabilistically) assigned. As a consequence, the approach simultaneously performs a dimensionality reduction and clustering, while enabling direct control over the organization of configurations in the latent space. We demonstrate the properties of this architecture by considering two model systems and molecular simulations of alanine dipeptide as well as a more challenging disordered peptide ensemble. In the following, $\mathrm{X} \in \mathbb{R}^n$ represents the n dimensional input. The latent variable in the bottleneck is represented by $z \in \mathbb{R}^d, d \leq n$.

6.3.1 One-dimensional 4-well potential

We first consider a single particle in one-dimension interacting with a 4-well external potential, which has been previously employed for testing methods associated with constructing MSMs [151, 132]. Figure 6.4(a) presents the potential, whose functional form and simulation details are given in Section 8.3.1. We employ a GMVAE with a latent space dimension of 1, which assesses the clustering performance of the architecture in the absence of any dimensionality reduction. The GMVAE was

trained with $k = 4$ according to the parameters in Table 6.2. Figure 6.4(b) presents the confusion matrix of the resulting model, which quantifies the probability that the model assigns a predicted label (x-axis) given the true label (y-axis). The true labels were determined using a coarse-grained representation of the system, where four metastable states are defined based on simple dividing surfaces, chosen as the maxima of the barriers between each potential well (dashed vertical lines in Figure 6.4(a)). The GMVAE assigns the state labels with 97% overall accuracy.

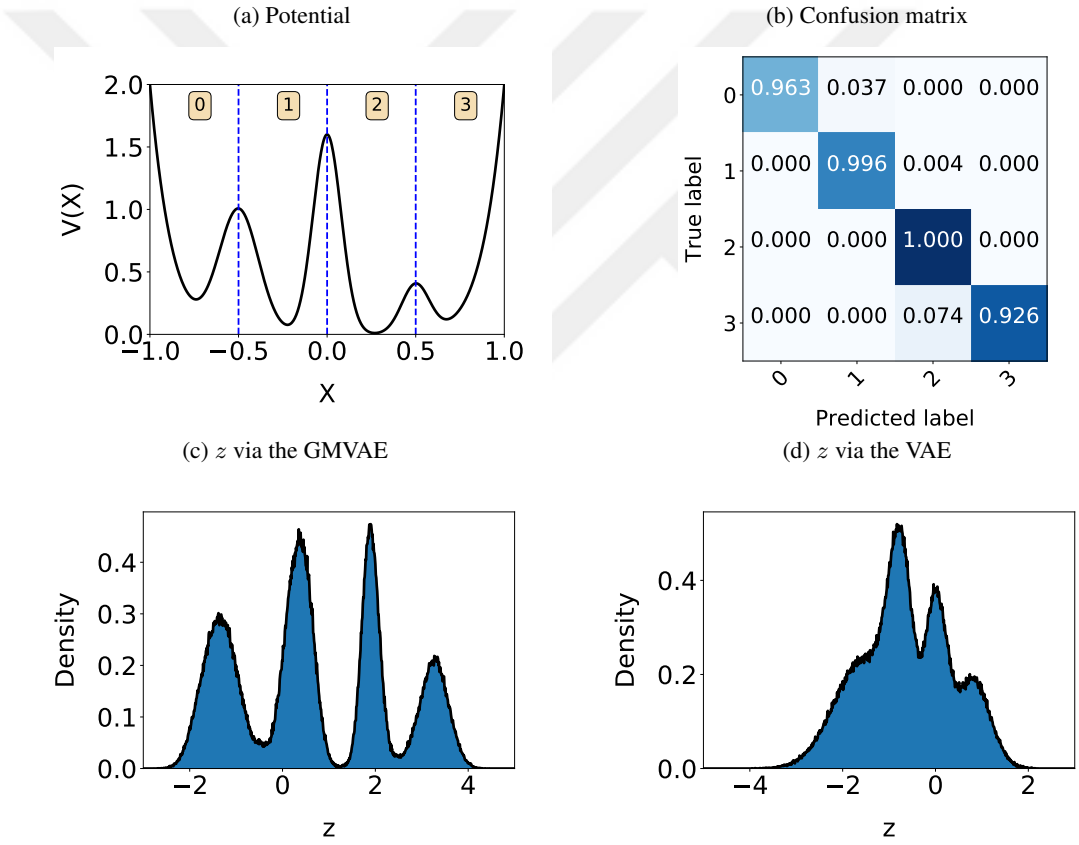


Figure 6.4: (a) 1D 4-well potential with the true labels. (b) Confusion matrix constructed with the true labels shown in (a) and the predicted labels obtained via the GMVAE. Population size increases from light to dark blue. Normalized histograms of the 1D latent variable via the (c) GMVAE and (d) VAE.

Figure 6.4(c) shows a normalized histogram of z values. Without dimensionality reduction, the GMVAE largely retains the description of the input space within

the latent dimension. As a consequence, the decoder is able to quite accurately reconstruct the input from the latent variable (See Figure 8.22). This behavior is in stark contrast to traditional VAEs, which employ a Gaussian prior to represent the latent space distribution. As a result, anti-clustering effects can arise, leading to highly overlapping clusters of data in the reduced space. To demonstrate this effect, we constructed a traditional VAE for the present example. Figure 6.4(d) presents the corresponding normalized histogram of z values. In this case, even without a reduction in dimension, significant information is lost due to the constraint of the assumed prior distribution.

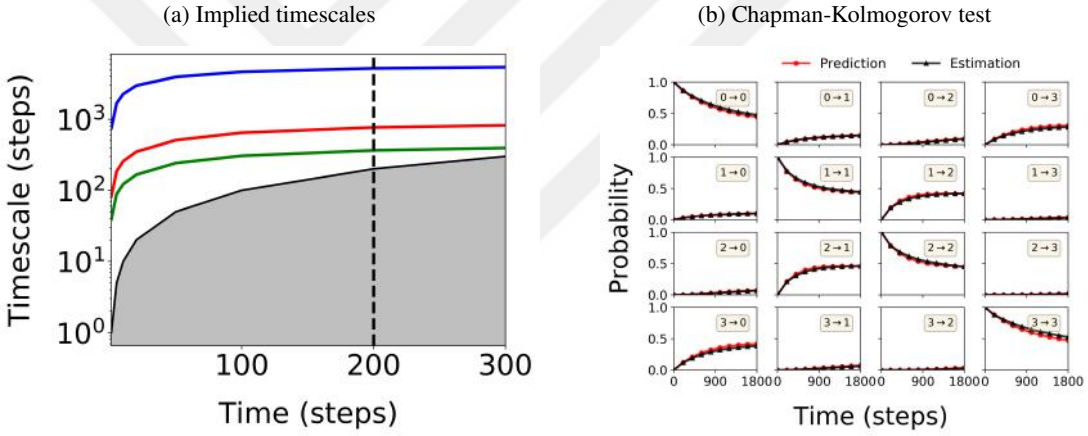


Figure 6.5: Markovianity check of the kinetic model built for 1D 4-well potential system. The MSM was constructed directly using the cluster labels obtained from the GMVAE. (a) Implied timescale test. (b) Chapman-Kolmogorov test (at lag = 200 steps).

To further characterize the quality of the GMVAE clustering, we constructed an MSM from the trajectories of the predicted cluster IDs. Figure 6.5(a) presents the standard implied timescale test, which assesses the convergence of the characteristic timescales with increasing lag time parameter τ. Convergence indicates that the simulation dynamics, within the discrete-state representation, can be described within a Markovian approximation. The gray area indicates timescales that cannot be resolved by the model, since they are faster than the chosen lag time. From

the test, the MSM with $\tau = 200$ was chosen for further analysis. The accuracy of this model was assessed with the Chapman-Kolmogorov test, which compares the simulated and predicted decay of probability from a chosen set of metastable states. Figure 6.5(b) demonstrates that the predicted "cluster dynamics" accurately represent the long timescale kinetic properties of the underlying simulation trajectory.

6.3.2 Müller-Brown potential

To assess both the dimensionality reduction and clustering performance of the GMVAE approach, we next consider a single Brownian particle in two dimensions interacting with an external Müller-Brown potential. The trajectory data was generated as the procedure suggested in [131] with the standard parameters [152] (see Section 8.3.2 for more details). As depicted in Figure 6.6(a), the resulting FEL contains two deep minima along with a less stable intermediate state. We employ a GMVAE that is trained with a latent space dimension of 1 and with $k = 5$, according to the parameters in Table 6.2.

Despite employing $k = 5$, the resulting GMVAE model identified only 3 states with non-zero membership probabilities. Thus, somewhat remarkably, the GMVAE architecture was able to identify the inherent organization of the input data in the high-dimensional space, independent of the hyperparameter k. Figure 6.6(b) shows the identified clusters. We define the true cluster labels in this case using linear dividing surfaces, as shown in Figure 8.23(a). Figure 6.6(c) presents the confusion matrix from the GMVAE model with respect to these defined labels. Although it appears that there are errors in assigning state 1, this error is sensitively dependent on the precise definition of the true label dividing surfaces. Moreover, the overall classification accuracy is actually 99%, since state 1 corresponds to a very rarely sampled intermediate state. The model also demonstrates relatively high reconstruction accuracy (See Figures 8.23(b) and 8.23(c)). Figures 6.6(d) and 6.6(e) present normalized histograms of z values obtained from the GMVAE model and a traditional VAE model trained on the same data, respectively. The low-dimensional representations obtained from the GMVAE clearly demonstrate a

better separation of metastable states. Additionally, the ability of the GMVAE to learn a nonlinear manifold is demonstrated in Figure 8.24, with respect to the linear embedding obtained using time-lagged independent component analysis (TICA).

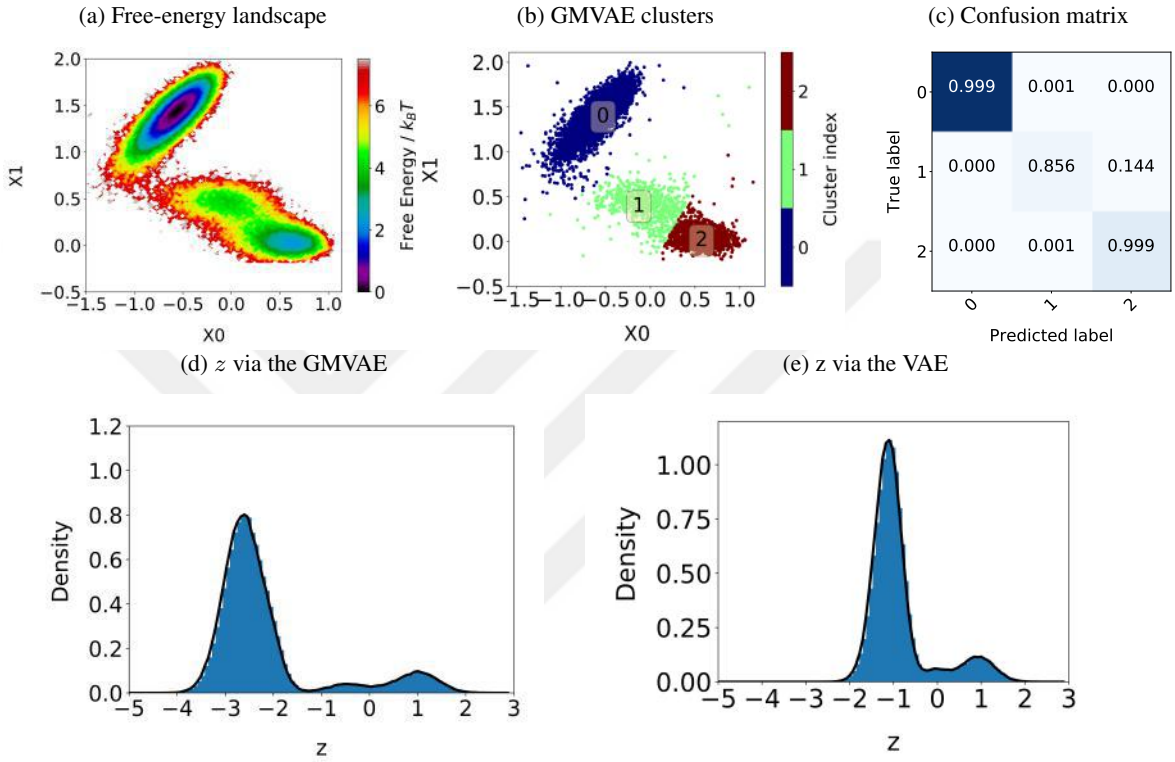


Figure 6.6: 2D Müller-Brown potential. (a) Free-energy landscape. (b) Clusters obtained from the GMVAE. (c) Confusion matrix with the true labels determined with linear dividing surfaces (Figure 8.23(a)) and predicted labels obtained via the GMVAE. Population size increases from light to dark blue. Normalized histograms of the 1D latent variable via the (d) GMVAE (e) VAE.

To further characterize the quality of the GMVAE clustering, we again constructed an MSM from the trajectories of the predicted cluster IDs. The implied timescale test (Figure 6.7(a)) shows two dominant processes. The MSM with $\tau = 10$ was chosen for further analysis. Figure 6.7(b) presents the Chapman-Kolmogorov test, which further verifies the accuracy of the GMVAE embedding.

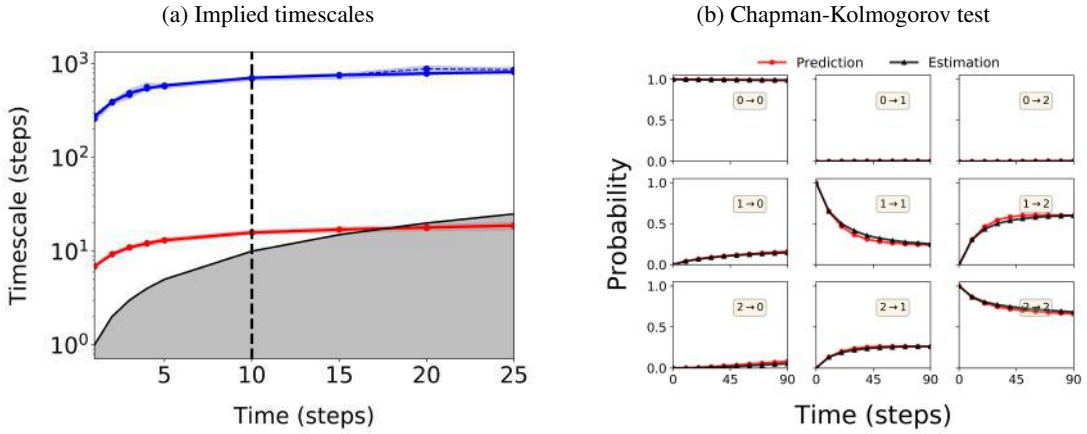


Figure 6.7: Markovianity check of the MSM built for 2D Müller-Brown potential via the GMVAE. (a) Implied timescales. (b) Chapman-Kolmogorov test (at lag=10 steps).

6.3.3 Alanine dipeptide

Alanine dipeptide is a representative model system for the characterization of conformational dynamics. Previous work [153, 134, 154, 130, 155] has shown that the (ϕ, ψ) backbone dihedral angles act as ideal collective variables for describing the metastable configurational basins and associated transition kinetics, making it an excellent system for testing the GMVAE framework within a more realistic molecular simulation context. Since in general the optimal set of input features is unknown a priori, we use this example to test the ability of the GMVAE to identify the proper collective variables from a larger set of input features. More specifically, we consider as input features both the normalized pairwise distances between heavy atoms as well as the (ϕ, ψ) dihedral angles (obtained from [156]). The pairwise distances were pre-processed using a kurtosis filter (with the threshold value of 0.03, see Figure 8.25 for more detail), to reduce the input dimension by removing the low-variance features. The dihedral angles were pre-processed by applying sin and cos transformations in order to account for periodicity [157]. Figure 6.8(a) shows the FEL in the backbone dihedral angle space, with four labeled metastable basins corresponding to α_{R}, α_{L}, β, P_{II}, and γ conformations [158]. The gray lines are drawn

for reference and do not represent any sort of optimal dividing surface.

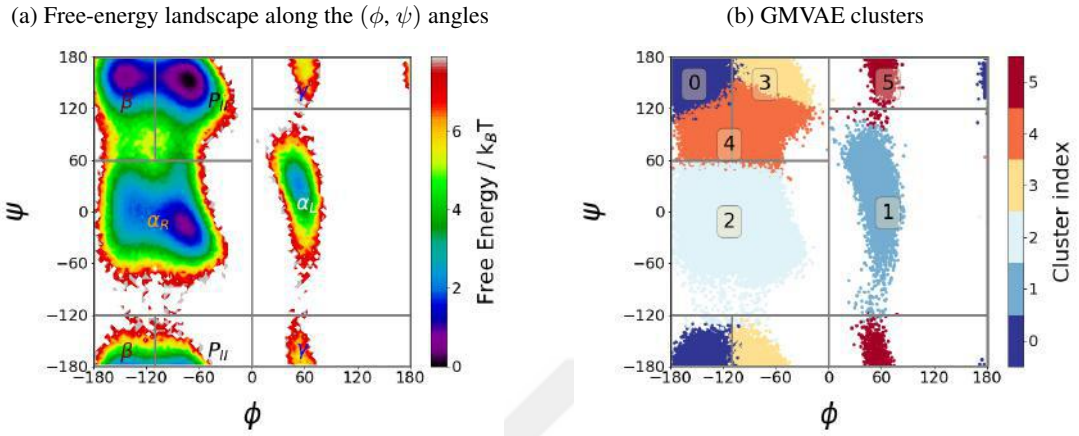


Figure 6.8: (a) Free-energy landscape of alanine dipeptide. (b) GMVAE clusters on the Ramachandran plot.

Figure 6.9(a) presents the two-dimensional embedding found using the GMVAE, and Figure 6.9(b) shows the simultaneously-obtained 6 clusters (indexed from 0 to 5) as a part of the GMVAE algorithm. The GMVAE again obtains a FEL that better separates clusters of conformations, relative to a standard VAE (Figure 8.29). The distribution of these clusters on the Ramachandran plot (Figure 6.8(b)) already strongly indicates their suitability for a kinetic analysis. The GMVAE clustering distinguishes all 5 of the metastable states, as well as a transition region between the α_{R} and β states (cluster 4). An MSM was again constructed from the coarse GMVAE cluster assignments. The implied timescale and Chapman-Kolmogorov tests are presented in Figure 6.10, demonstrating the accuracy of this kinetic model.

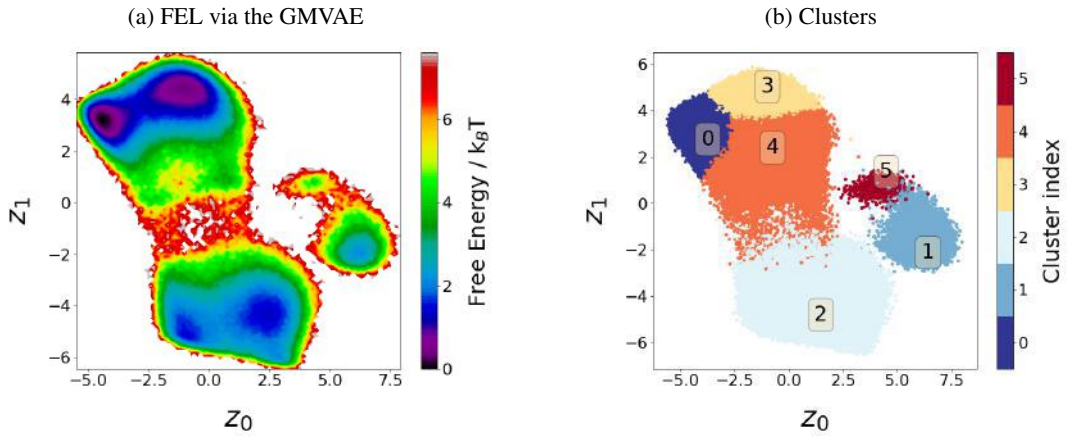


Figure 6.9: (a) FEL obtained for the alanine dipeptide by the GMVAE. The GMVAE clusters on the (b) GMVAE landscape.

We found in this example that, unlike the toy systems, the clustering obtained using the GMVAE did not appear to be completely robust. In particular, the precise clustering probabilities depend on the random effects of the training procedure (e.g., random weight initialization and the random shuffling of the input data). This issue was most pronounced for the lowest populated state, whose probability differs from the other states by two orders of magnitude (Figure 8.26(b)). As a consequence, the γ state was not always sufficiently separated from the α_{L} state, resulting in a loss of one of the resolved kinetic processes (although the accuracy of the MSM remained intact, see Figure 8.28). Despite this issue, the obtained FEL appeared rather robust with respect to changes in the random factors during training. We observed a much more robust clustering for all other applications considered.

6.3.4 $AAQAA_3$ peptide - I

As a more challenging test, we consider simulation trajectories of the capped helix forming peptide AC-$(AAQAA)_3$-NH2, which is a representative system for investigating helix-coil transitions. We employ a coarse-grained model [159], which describes the dominant attractive interactions, e.g., hydrogen bonding and effective hydrophobic interactions between side chains, with simple potentials between

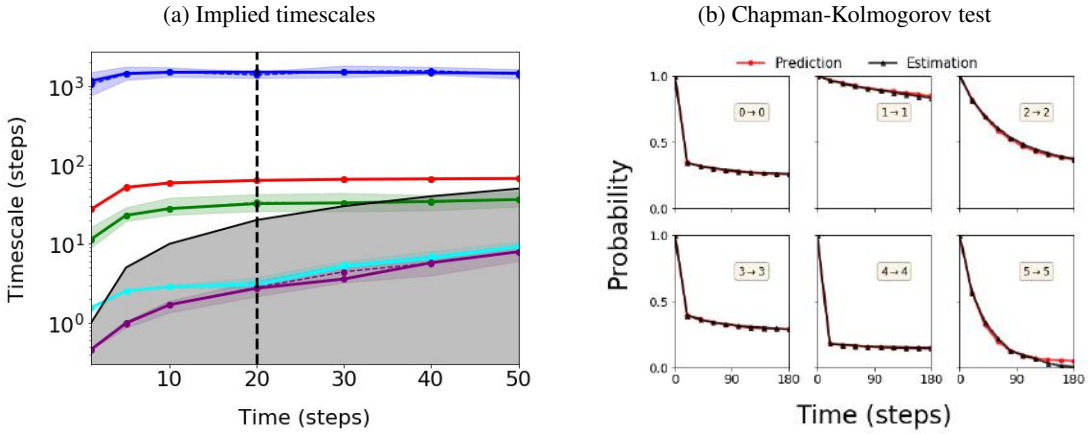


Figure 6.10: Markovianity check of the MSM built for alanine dipeptide via the GMVAE. (a) Implied timescales. (b) Chapman-Kolmogorov test (at lag=20 steps).

the C_α and C_β atoms. These interactions are the minimum required to sample the proper range of structures, (i.e., helix, coil, and hairpin-like). This model also represents excluded volume effects in near-atomic detail, which was demonstrated to be important for accurately characterizing the helix-coil kinetics. Here we employ a parametrization of the model that most accurately reproduces the experimental cooperativity of the helix-coil transition for $AAQAA_3$. As a result, hairpin-like structures appear to have relatively low metastability (similar to the intermediate state in the Müller-Brown example, and the γ state in alanine dipeptide), as we discuss further below. The model and simulation protocol are discussed further in the Section 8.3, and also in [159, 160]. The considered simulation trajectories correspond to a disordered ensemble of peptide configurations, representing a stringent test for dimensionality and clustering methods [161].

Similar to alanine dipeptide, the set of sin and cos augmented (ϕ, ψ) dihedral angles along the peptide backbone were used as conformational descriptors. Thus, the input dimension is 60 for the 15-residue $AAQAA_3$ peptide. We chose to consider only a latent space dimension of 2, given that the ultimate goal of dimensionality reduction is often to reduce the high-dimensional description to something that is easily visualizable. Unlike the simple model systems above, the number of clusters,

k, is completely unclear a priori. In fact, we expect that this ensemble to have a hierarchical structure, such that differing number of clusters may be appropriate depending on the chosen level of resolution. While we initially considered the GMVAE with varying number of clusters, we found that the number of "non-zero clusters" (i.e., clusters with a significant probability of configuration assignment) was extremely insensitive to this choice, as discussed below. The GMVAE was trained according to the parameters in Table 6.2. Also in contrast to the previous examples, there is no definitive reference kinetic model with corresponding known metastable states. Instead, the analysis below assesses the GMVAE embedding and clustering (in terms of both statics and kinetics) with respect to the landscapes obtained using a standard VAE and also following the standard MSM workflow (i.e., TICA [123, 145], see Section 6.2.4 for more details). Panels (a) and (b) of

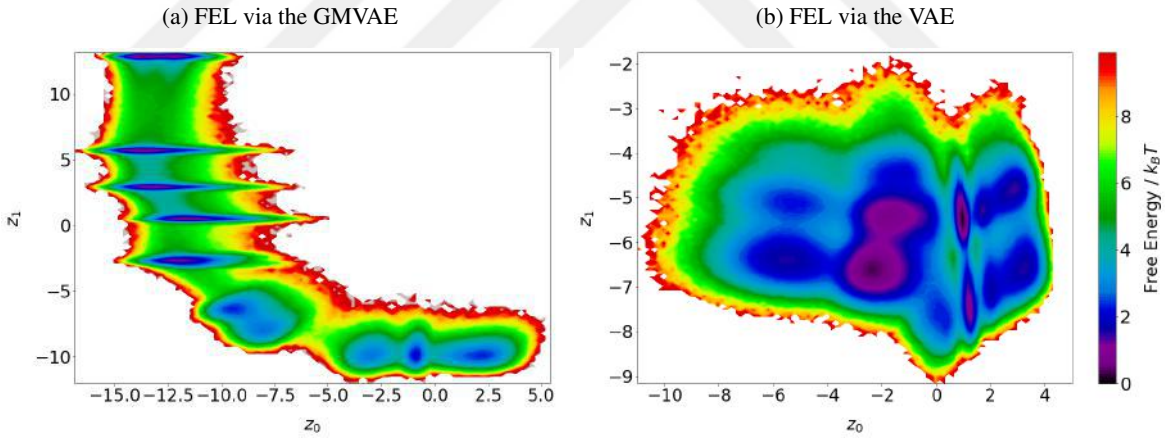


Figure 6.11: Free-energy landscapes of $AAQAA_3$ - I peptide obtained by (a) the GMVAE, and (b) the VAE.

Figure 6.11 show the FELs obtained using the GMVAE and the traditional VAE, respectively. As in the model systems, the GMVAE method results in a latent space description with highly separated clusters, while the traditional VAE yields more overlapping states. The two-dimensional TICA landscape (Figure 8.40) also separates a number of clearly distinct states, although there are large diffuse regions with relatively low free-energy values. The clusters obtained via the GMVAE are

shown in Figure 6.12(a). Despite employing $k = 10$ and obtaining a landscape that appears to have approximately 10 distinct basins, only 7 states (labeled $0, 1, \ldots 6$) were assigned non-zero membership probabilities (see Figure 8.30). Since standard metrics for analyzing peptide configurations do not yield a clear organization of the ensemble into a small number of metastable states, the distribution of these quantities are expected to be highly overlapping, even for a good clustering of the input data. Thus, to more easily visualize the characteristics of the GMVAE clusters, we applied a thresholding scheme, which removes configurations without a membership probability greater than 0.95 (see Section 6.2.3.1 for details and Figure 8.31 for cluster populations). Figure 6.12(b) shows 5 representative structures closest to the cluster centers. We stress that these images are intended to give the reader a rough idea of the types of structures contained in each cluster, but do not characterize the variance of structures within the clusters. This is a disordered ensemble and each cluster necessarily contains a diversity of structures. Nevertheless, Figure 6.12(b) indicates that the GMVAE successfully distinguishes between distinct secondary structures within the simulation data.

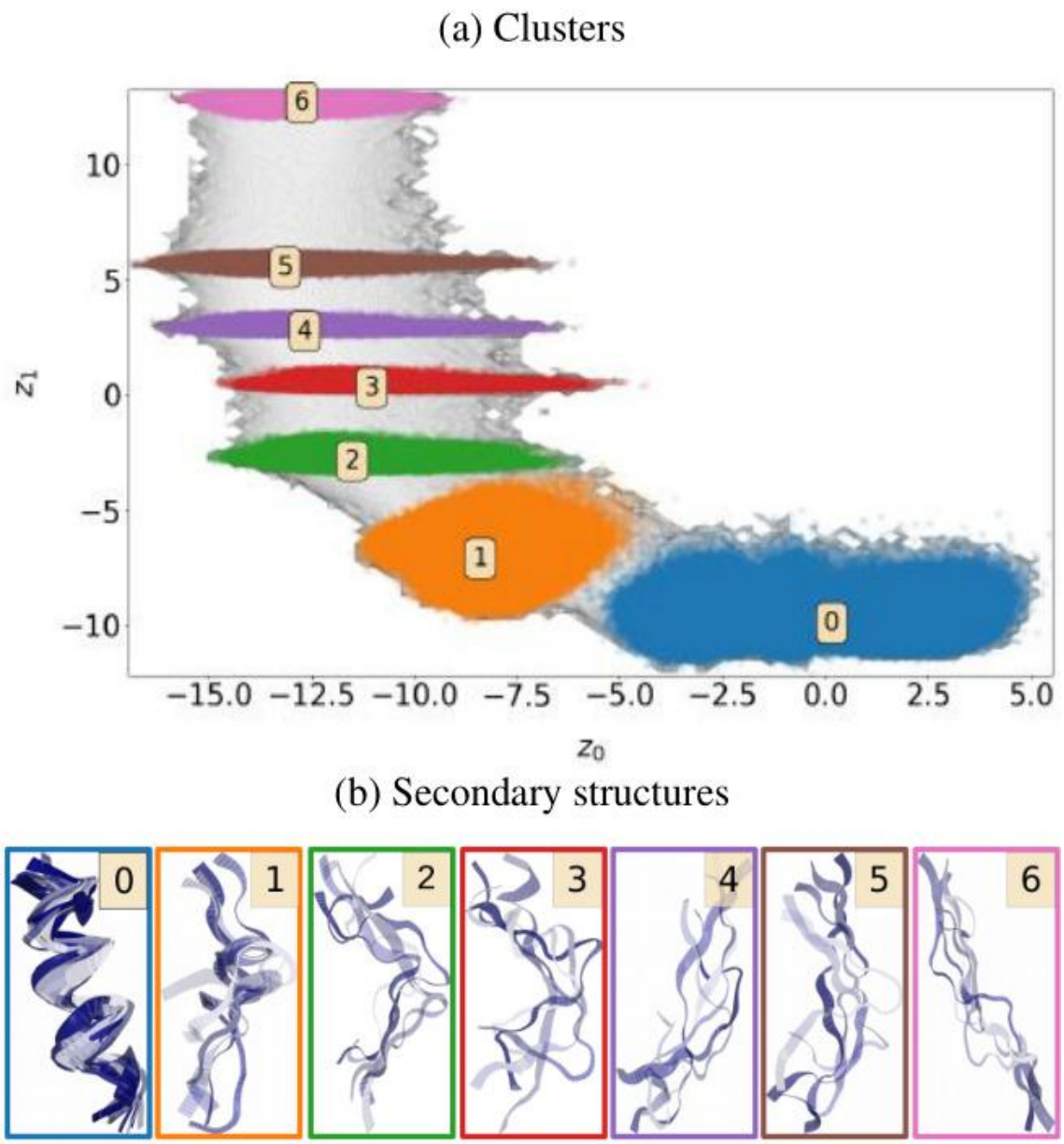


Figure 6.12: (a) The clusters obtained for the $AAQAA_3$ peptide - I by the GMVAE after thresholding. (b) The secondary structures closest to the cluster centers.

To characterize the structural properties of the clusters quantitatively, we calculated the distribution of the average fraction of helical segments, $\langle f_h \rangle$. Figure 6.13(a) presents a heat map of $\langle f_h \rangle$ in the latent space. High $\langle f_h \rangle$ values (represented by blue) indicate the presence of helix and helix-like structures, whereas the lower values point to either hairpin- or coil-like secondary structures. There is an apparent

trend of decreasing average helical content from the lower-right to upper-left regions of the latent space (i.e., from cluster 0 to 6). The VAE and TICA landscapes demonstrate similar trends (Figures 8.44(b) and 8.40(b), respectively), although the VAE does not characterize partially-helical structures as clearly as the GMVAE. Figure 8.32 presents the intra-cluster distributions of $\langle f_h \rangle$, which can be used to assess the quality of the clustering (relative to an alternative clustering). We expect that an optimal clustering will result in tight, unimodal $\langle f_h \rangle$ distributions. The GMVAE clustering yields seemingly good distributions for the most and least helical clusters, while the partially-helical clusters appear broader and somewhat bimodal. For comparison, we consider three alternative clusterings obtained by performing a k-means clustering on a given landscape followed by the PCCA$^+$ dynamical coarse-graining method [146] to define a set of metastable states (see Section 6.2.4 for more details): (i) an alternative clustering of the GMVAE landscape (Figure 8.37), (ii) a clustering on the VAE landscape (Figure 8.45), and (iii) a clustering on the TICA landscape (Figure 8.41). The alternative clustering scheme on the GMVAE landscape, (i), does not improve the intra-cluster distributions of $\langle f_h \rangle$, demonstrating that the GMVAE clustering is reasonable, given the GMVAE embedding. Similar results were obtained from the VAE clustering, with slightly broader distributions for the most and least helical states. The TICA clustering resulted in somewhat improved distributions, in the sense that they appear to be mostly unimodal, although some of the distributions appear to be slightly broader.

Figure 6.13(b) shows the $\text{dRMSD}_{\text{hel}}$ values of the projections, where the helicity increases as the $\text{dRMSD}_{\text{hel}}$ values decrease. These results are in agreement with the $\langle f_h \rangle$ analysis: as the cluster index increases from 0 to 6, the conformations tend to be more extended. The Supporting information in Section 8.3 (Figures 8.33 and 8.34) contains additional characterization of the static properties of the clusters, which further validate the GMVAE embedding and clustering as a reasonable partitioning of the conformational landscape.

We also characterized the average fraction of helical segments on the N- and C-terminus sides of the peptide: $\langle h_N \rangle$ and $\langle h_C \rangle$, respectively (see Section 6.2.5

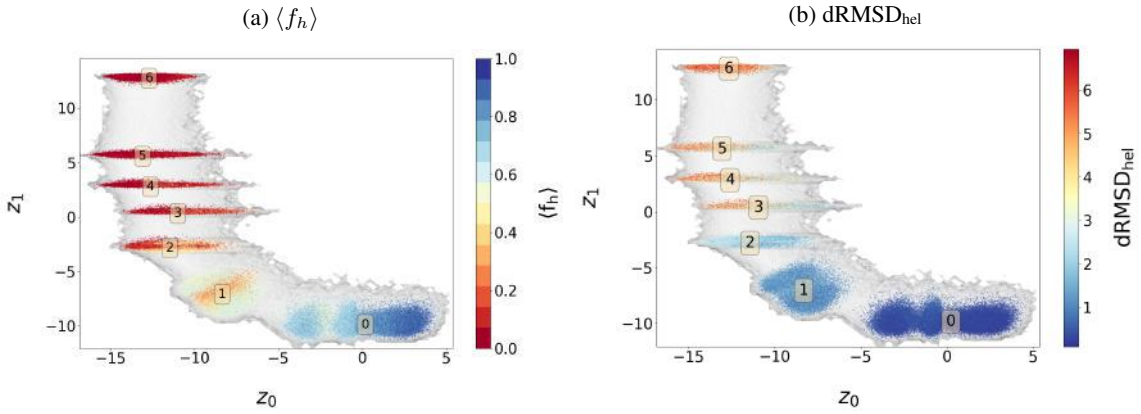


Figure 6.13: $AAQAA_3$ - I. (a) Average helical fraction, $\langle f_h \rangle$, analysis. Colors represents the $\langle f_h \rangle$ values of the corresponding projected data obtained from the GMVAE. (b) $dRMSD_{hel}$ analysis.

for more details). Figure 6.14 presents the difference of these quantities, $\langle h_N \rangle - \langle h_C \rangle$, plotted along the GMVAE embedding. Positive values (represented by blue) indicate conformations that contain helical structure on the N-terminus side of the peptide without helical structure on the C-terminus side. Conversely, negative values (represented by red) indicate conformations that contain helical structure on the C-terminus side of the peptide without helical structure on the N-terminus side. Values close to zero correspond to either fully helical or non-helical structures. Although the GMVAE embedding and clustering separate the most distinct structures in the ensemble (coils and full-helicies), some of the clusters (0, 1, 2) encompass partially-helical conformations on both sides of the peptide (see also Figure 8.36). This is not ideal since kinetic barriers within a cluster will negatively impact the accuracy of a kinetic characterization at the cluster level. However, it appears that this issue may have more to do with the clustering than the embedding itself, since blue- and red-labeled structures appear to be reasonably separated on the landscape.

Similar to the other examples above, we also constructed an MSM directly from the discretized trajectories of GMVAE cluster indices. Although thresholding was applied in the results presented here (practically similar to coring methods for constructing kinetic models [162]), we found that this procedure had negligible effect

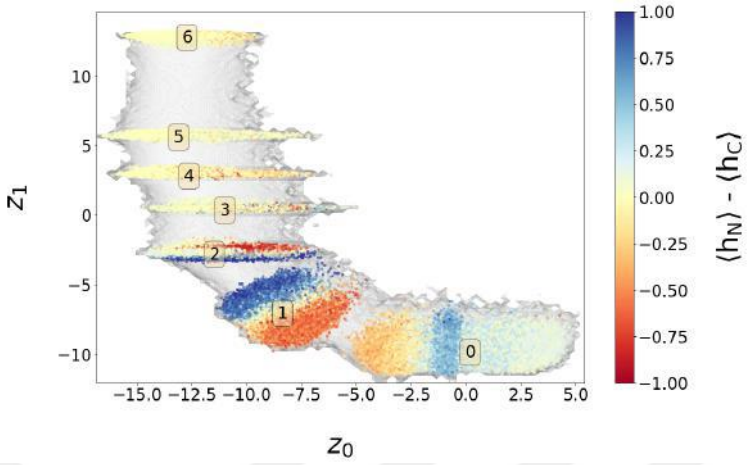


Figure 6.14: Analysis of partially-helical conformations for $AAQAA_3$ - I. Projections are colored according to $\langle h_N \rangle - \langle h_C \rangle$ values.

on the accuracy of the resulting MSM. As shown in Figure 8.35, the MSM constructed from the GMVAE clustering displayed significant errors in describing, e.g., the decay of probability out of the helix state. Perhaps this is not so surprising, since coarse-grained MSMs are often only used as a qualitative analysis tool, while higher-resolution kinetic models that characterize configuration space with many microstates are used for quantitative reproduction of simulation kinetics. Thus, to more carefully assess the GMVAE embedding and to more easily compare to the VAE and TICA results, we constructed a higher-resolution MSM by performing k-means to define microstates on the landscape (Figure 8.37). Although the resulting model demonstrates improved accuracy according to the Chapman-Kolmogorov test, the probability decay out of the metastable states occurs on a fast timescale relative to the chosen lag time. This may be indicative of poorly defined dividing surfaces between metastable states. The kinetic models constructed from the VAE and TICA landscapes (Figures 8.45 and 8.41, respectively) demonstrate similar quickly decaying probabilities. Although coring procedures could be applied to attempt to fix this problem, it indicates that there are fundamental limitations of all of these landscapes in terms of characterizing the long timescale simulation kinetics. There are several possible reasons for these difficulties, including (i) the limitation

of our embeddings to two dimensions, (ii) the limitation of the chosen input features in characterizing kinetically-distinct structures, (iii) the presence of many low-lying barriers along the potential energy landscape of this disordered ensemble, and (iv) the poor sampling of relatively rare transitions to the full helix conformation. We partially address items (ii) and (iv) in the next section; however, a detailed investigation of these issues is beyond the scope of this initial study of the performance of the GMVAE, and is left for future work.

6.3.5 $AAQAA_3$ peptide - II

To investigate the impact of the low sampling of helical structures on the GMVAE embedding, as in the $AAQAA_3$ - I simulations presented above, we also considered a second set of simulations which primarily samples helical- and hairpin-like structures, while only rarely sampling fully-coil structures. (Please see the Supporting Information (Section 8.3) for more details about the differences between the two sets of simulations). In addition to the dihedral angles, normalized pairwise distances between residues that are more than 3 residues apart were included as input features. Figure 6.15 presents the obtained GMVAE FEL (panel (a)), the corresponding clustering of 6 metastable states (panel (b)), and overlays of five structures that are closest to the cluster centers (panel (c)). The GMVAE embedding demonstrates significant separation of metastable states, relative to the landscape obtained with a standard VAE (Figure 8.58(a)). Similar to the previous ensemble ($AAQAA_3$ - I), Figure 6.16 shows the separation of structures according to $\langle f_h \rangle$ (panel (a)), and $dRMSD_{hel}$ (panel (b)). The VAE and TICA landscapes demonstrate similar trends (Figures 8.58 and 8.54, respectively). The intra-cluster $\langle f_h \rangle$ distributions are shown in Figure 8.49. The majority of the fully-helical structures are in cluster 3 and 5, while clusters 0, 1, 2 and 4 contain hairpin-like structures as well as partial helicies. The coil structures are gathered in the bottom-most part of the landscape (in cluster 4), though not separated as a distinct cluster by the GMVAE. The distributions are broader and less unimodal than those determined from the previous set of simulations, although these can be somewhat improved with the alternative

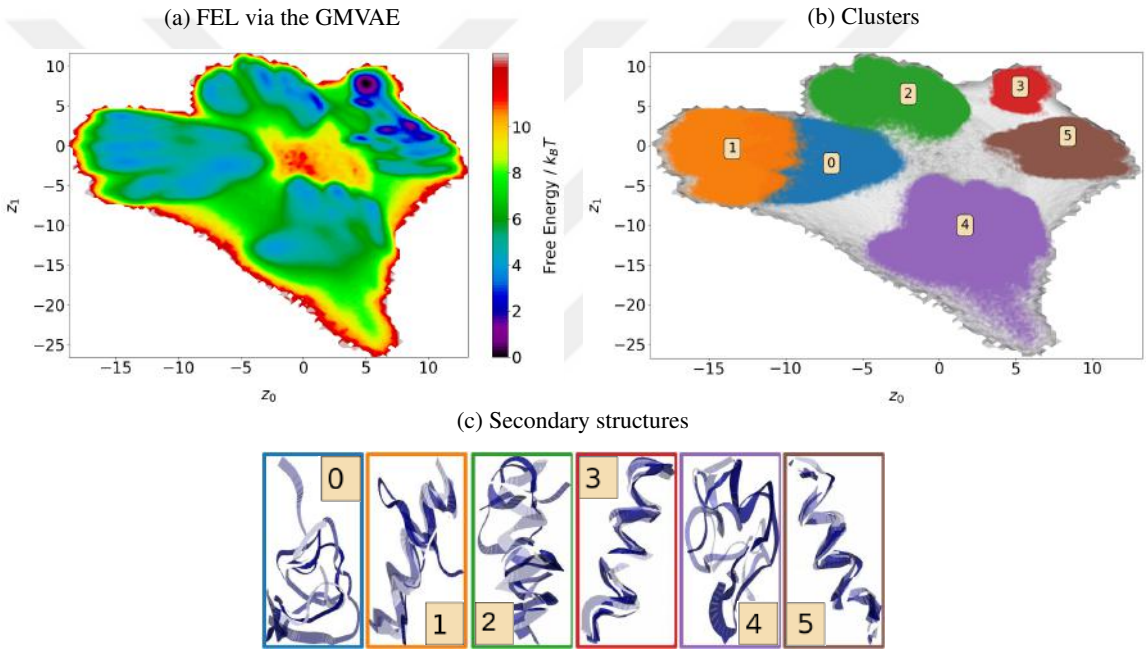


Figure 6.15: The GMVAE results for $AAQAA_3$ peptide - II. (a) Free-energy landscape. (b) The clusters obtained after thresholding. (c) The secondary structures closest to the cluster centers.

clustering scheme on the GMVAE landscape (Figure 8.53). Similar results are also obtained from the VAE and TICA landscapes (Figures 8.61 and 8.57, respectively). Figure 6.17 presents the characterization of the N- and C-terminus, partially-helical

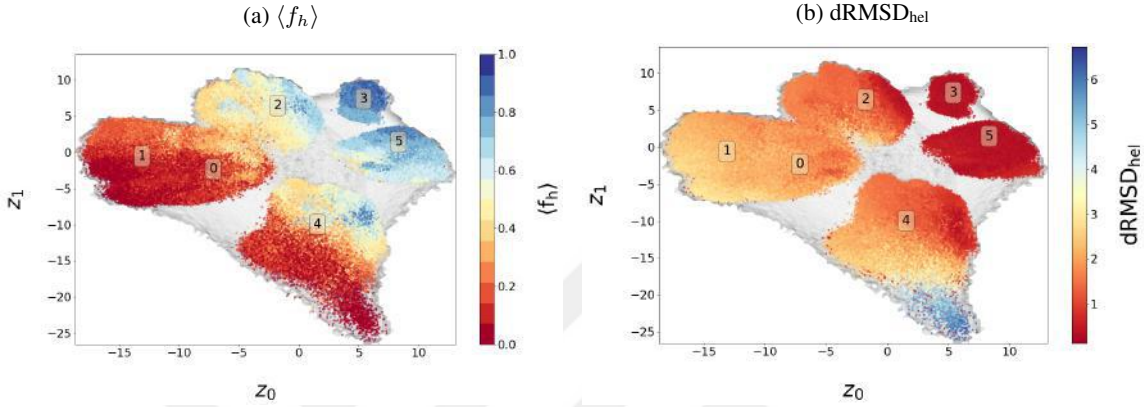


Figure 6.16: Projections for the AAQAA$_3$ peptide - II. (a) $\langle f_h \rangle$, (b) dRMSD$_{\text{hel}}$.

conformations. In contrast to the AAQAA$_3$ - I embedding, the GMVAE embedding and clustering for AAQAA$_3$ - II more clearly separates the distinct types of structures. It appears that this difference may be due to the increased sampling of helical structures in AAQAA$_3$ - II, although the inclusion of pairwise distances as additional input features may also have played a role. N- and C-terminus partially-helical structures are mostly located in clusters 4 and 2, respectively, while both types of structures can be found to a lesser extent in cluster 5. Although the VAE and TICA landscapes also appear to largely distinguish between distinct partially-helical structures (Figures 8.58 and 8.54, respectively), the GMVAE landscape provides a significantly better clustering of these two distinct sets of conformations.

Despite the improved description of partially-helical structures, the MSM constructed directly from the GMVAE clustering for AAQAA$_3$ - II displayed similar discrepancies to the model built for AAQAA$_3$ - I (Figure 8.50). Moreover, the high-resolution MSMs constructed from the GMVAE, VAE, and TICA landscapes (Figures 8.51, 8.59, and 8.55, respectively) displayed very fast decay of probability out of the identified metastable states, as in the AAQAA$_3$ - I example.

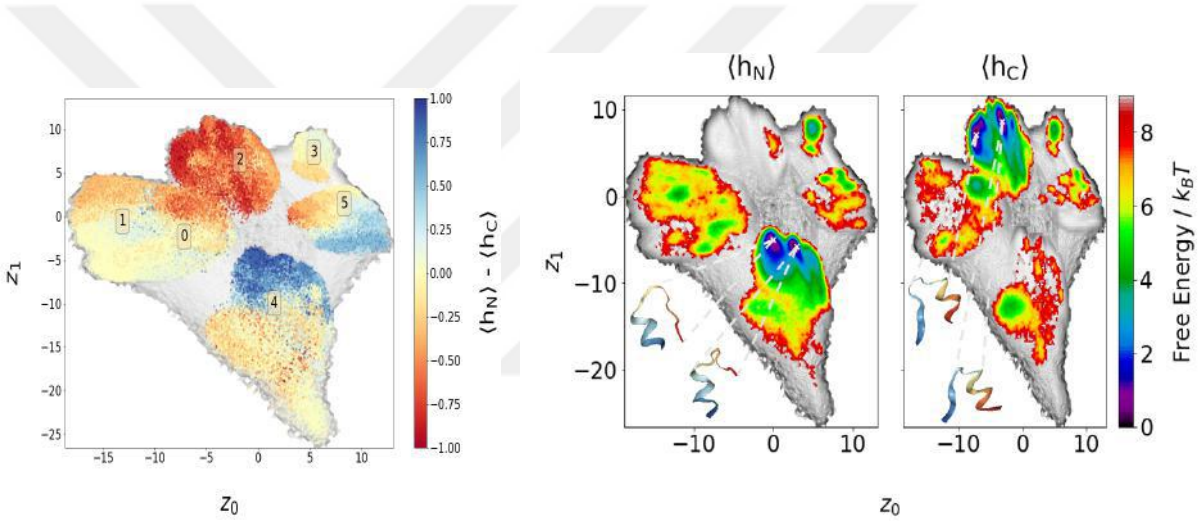


Figure 6.17: The N- and C-terminus end folding analysis for the $AAQAA_3$ peptide - II. (Left) The difference in the average values of the two-end foldings, $\langle h_N \rangle - \langle h_C \rangle$. (Right) Distribution of the N- (on the left, $\langle h_N \rangle \geq 0.8$) and C-end (on the right, $\langle h_N \rangle \leq -0.8$.)

Chapter 7

CONCLUSION

Chapter 2 gives an overview of the governing equation of motion for the protein dynamics.

Chapter 3 contains a short review of the methods used in approximating negative gradient of the potential energy function via linearization. We presented the nodal and node-branch formulations and showed that these formulations in fact constitutes a more general scheme, allowing the linearizations around not only equilibrium but also non-equilibrium points. We also showed that a reduced-size formulation can be derived only for the expansions around equilibrium points, which we also presented its equivalence to a commonly-used approach.

In **Chapter 4**, we presented PAC, a frequency domain technique for analyzing protein and ligand interaction dynamics that was inspired by an electronic circuit analysis scheme. We compared PAC with previously proposed techniques on several example proteins with ligand binding structures. PAC generalizes and subsumes the previous methods by incorporating the frequency of excitation as a key new parameter for dynamic analysis of proteins. We have proposed several new frequency dependent metrics for the characterization of the 3D complex response of system response that help interpret and extract useful information about the protein structure and dynamics. Finally, we should note that the PAC formalism is based on a dynamical model that is obtained by *linearizing* the nonlinear force field of the protein around a reference configuration, e.g., its native state. This enables the solution of the equations of motion directly in the frequency domain as captured by the PAC formalism, in an efficient manner, producing dynamic characterizations that would be very expensive to generate with time stepping based MD simulations. Thus, PAC enables high-throughput scanning studies.

In **Chapter 5**, we have investigated the phenomenon of allostery using well established tools of telecommunication systems, namely frequency decomposed SNR and channel capacity based analyses. To that end, we analyzed the displacements due to both noise forces and external excitations—capturing the effect of the ligand—separately. We proposed two related analysis schemes, termed as *per residue scan* and *binding pocket excitation*, in order to identify the residue pairs that are likely to interact allosterically, and the residues that are affected in a significant manner by a particular ligand-binding event, respectively. The frequency domain representations employed by the proposed methods lead to an alternate view into allostery, by emphasizing the effect of perturbation frequency in the SNR response. We have shown that the response of some of the residues exhibit a resonance at specific, characteristic frequencies. Thus, a full spectral analysis of the responses to perturbations leads to the speculative but potentially significant conclusion that the key mechanism underlying allostery is robust signal transmission despite noise at specific frequencies. The frequency-decomposed approach further allows the analysis of equilibrium fluctuations in the absence of the ligand. We proposed an alternative technique to compute mean square fluctuations in a frequency-decomposed manner.

In **Chapter 6**, we propose a Gaussian mixture model as the prior distribution in the latent space of a variational autoencoder, to explicitly enforce multi-basin structure of an ideal free-energy landscape that characterizes basins that are well-separated by the largest barriers along the higher-dimensional potential energy landscape. The performance of the Gaussian mixture variational autoencoder (GMVAE) was illustrated on two standard toy-model systems and on the standard benchmark alanine dipeptide, as well as on a challenging 15-residue-long disordered peptide. For each example, the GMVAE circumvents the aggregation of points in the latent space characteristic of traditional variational autoencoders. Instead, samples that are structurally distinct are clearly separated, leading to a latent space that displays apparent metastable basins and barriers. The GMVAE introduces a categorical variable that probabilistically assigns samples to a set of underlying clusters, each of which is Gaussian distributed. Thus, the approach combines the commonly distinct

tasks of dimensionality reduction and clustering into a unified framework. In the context of static equilibrium properties, the incorporation of the Gaussian mixture model as a prior distribution on the latent space closely links our physical intuition about ideal free-energy landscapes, resulting in an inherently more interpretable latent space. Our results show encouraging performance when constructing kinetic models from the learned representations, serving as an independent validation of the procedure.

Chapter 8 presents additional results to the chapters and provides further details.

The findings of this study point to a number of important improvements for future studies. A single-point force application may not always be able to adequately model ligand binding. In such cases, applying forces simultaneously to all of the residues that interact with the ligand may serve as a better model for the ligand binding event. However, this would require a more complicated force configuration determination procedure in cases where no MD simulation is possible. Furthermore, for proteins with homogeneous structure and smaller-scale functional motions, ligand binding may be better modeled as a dynamic event with a time-varying, multi-frequency force perturbation. We are working on the development of a scheme for determining the parameters of multi-point, dynamic (time-varying) force perturbations that could serve as a better model for ligand binding. In our future work, we plan to extend PAC so that one can perform frequency domain analyses of protein and ligand binding dynamics with multi-frequency excitations. Additionally, the full potential of the proposed SNR and channel capacity based analysis methods has not been fully explored yet. More detailed analyses are required to determine the full benefits and limitations. Regarding the representation learning, we argue that incorporating physical constraints into the architecture helps to construct an interpretable model for the kinetics, even when kinetic information is not used for learning the representation. Although higher-resolution MSMs constructed directly from the GMVAE landscape demonstrated an improved description of the simulation kinetics, the resulting model for the disordered ensemble was un-

able to resolve all but the longest timescale processes. An MSM constructed from the TICA landscape demonstrated a slight improvement over this model, with respect to the CK test, but also exhibited a very fast decay of probabilities out of the identified metastable states, indicating a significant limitation in the time resolution of the model. These issues highlight the difficulty of characterizing such disordered ensembles, and motivate further investigation into the various possible causes. For example, comparisons of two distinct peptide ensembles clarified the role that sampling can play in distinguishing distinct partially-helical structures on the GMVAE landscape. It remains unclear to what extent the restriction of our embeddings to two dimensions or the choice of input features prevented the GMVAE (as well as the more standard methods considered) from better describing the simulation kinetics. Moreover, the presence of many low-lying barriers along the potential energy landscape of this disordered ensemble may cause fundamental challenges in obtaining a clear few-metastable-state characterization of the conformational landscape. Thus, we propose that, in conjunction with simpler test systems that clearly assess a method's performance, such examples are important for significant advancements in data-driven characterizations of molecular simulation trajectories. While we defer a more detailed investigation of these issues for future work, we highlight the promising performance of the GMVAE demonstrated through our results.

Chapter 8

APPENDIX

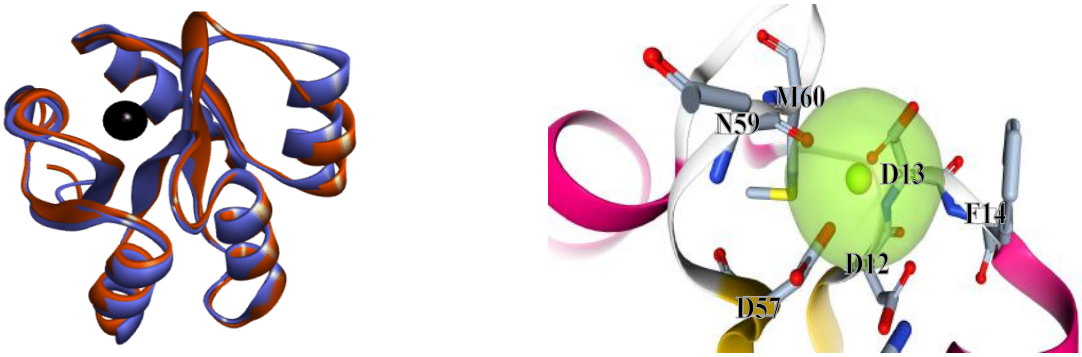
This chapter provides supplementary information for the previous chapters.

8.1 Supplementary information for Chapter 4: PAC: A frequency domain technique for analyzing protein dynamics

8.1.1 Chemotaxis signaling protein Y

CheY is a bacterial chemotaxis protein with 128 residues. Binding of Mg^{2+} cation to the active site results in a significant conformational change in the structure of the protein. Several residues are displaced at around 10 Å as a result of this conformational change [77]. The holo form of the protein is with PDB ID:1CHN (at a resolution of 1.76 Å), and the apo form is with PDB ID:3CHY (at a resolution of 1.66 Å). The residues are numbered from 2 to 129, and in the holo form ALA2 and ALA3 coordinates are missing. For our analysis, we include them using Modeller [78], however, for most of the calculations, we omit the first 2 residues. The ligand is in close interaction with residues ASP57, ASP13, and ASN59. Molecular dynamics simulations suggest that the $\beta_4 - \alpha_4$ loop (residues 86-92) in CheY acts allosterically as a gating element [22, 79]. We perform PAC analysis on CheY by applying single-point force excitations separately to two residues, one of the binding site residues, residue 57, and one of the residues of the $\beta_4 - \alpha_4$ loop, residue 91.

Displacement magnitudes, $|\Delta \boldsymbol{r}_{(i)}(\omega)|$, of C_α atoms are computed after the alignment of the ligand-bound form to the ligand-free form. After alignment, RMSD between the two structures is calculated as 1.6930 Å. The correlation coefficient between the experimental displacements and the computed ones (after alignment, with



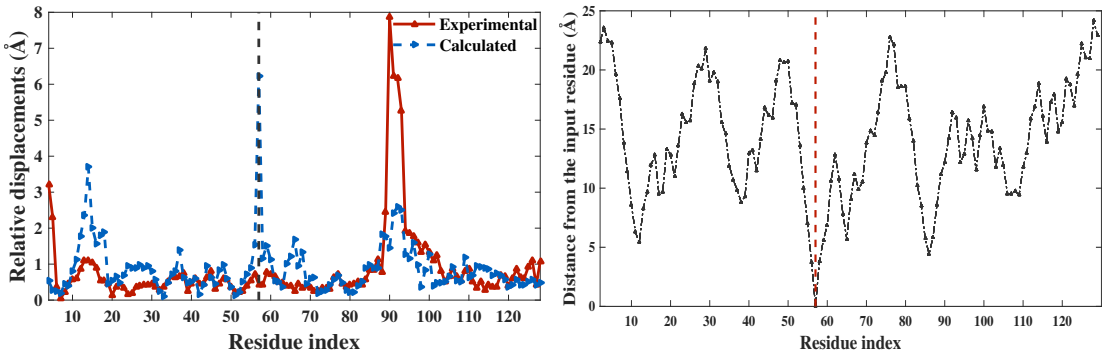
(a) Apo form (3CHY) and holo form (1CHN) of CheY are aligned, apo form is colored with blue and holo form is colored with orange, both in solid ribbon form. Ligand (Mg^{2+} ion) is shown as a black circle.

(b) Binding pocket of Mg^{2+} ion and interacting residues: D12, D13, F14, D57, N59, and M60. Plot was created using [75, 76].

Plots were created with [74].

Figure 8.1: Alignment of apo and holo forms and binding site of CheY

zero frequency excitation) is found to be 0.3496 if the force is applied to a binding pocket residue, residue 57, and 0.6942 if the force is applied to residue 91. The input residues are again selected from the PRS paper [22]. In Figure 8.2(a), and Figure 8.9(a), the red curves show the experimental displacements and the blue ones are for the calculated when the input residue is 57 and 91, respectively. Residue distances to these input residues are shown in Figure 8.2(b) and Figure 8.9(b). Figure 8.3(a) shows the 3D displacements and Figures 8.3(b) and (c) are for side views. Displacement values above 10 Å is not shown in the figures. Force frequency definitely affects the displacements in a residue-specific manner. Figure 8.3(d) shows the displacements for a set of selected residues 4, 14, 41, 56, 76, 93, 111. Displacement of residue 14 tends to decrease with frequency but at 0.1-0.25 THz, there is an increase. Residue 56 shows two different peaks at higher frequency values. Figure 8.3(e) shows the displacements for all of the residues at five selected frequencies in logarithmic scale.



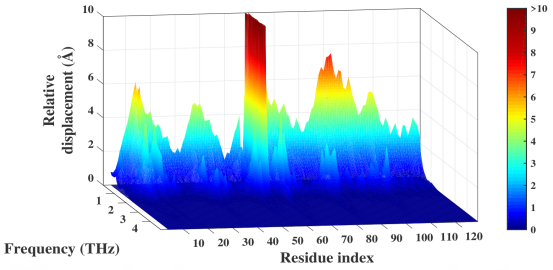
(a) Experimental and calculated displacements between apo and holo forms of CheY. Force excitation on residue 57.

(b) Residue distances to the input residue (57).

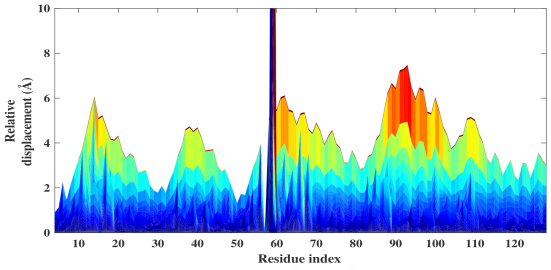
Figure 8.2: Displacement magnitudes of CheY residues in Å. Red curve shows experimental displacements and blue curve is calculated displacements at the best force magnitude and direction at zero frequency. Residues 10, 11, 54, 55, 56, 59, 63, 64, 84, 85, 86 are within 8 Å range of residue 57.

Kinetic energy, $E_i(\omega)$. In accordance with the same *critical* residue selection criteria that was defined in Section 4.5, the critical residues for CheY are 11, 12, 17, 56, 57, 66, 68 when input residue is 57, and residues 87-97, 110 when input residue is 91 as shown in Figures 8.4(a) and 8.11(a). At low frequencies, energy is highly correlated with the inverse of the inter-residue distances (i.e., proximity) when force is applied to residue 91, however for excitations applied on residue 57, energy becomes correlated with the proximity at high frequencies. The force application site not only determines how much ligand binding is mimicked but also the frequency dependent internal energy distribution within the protein.

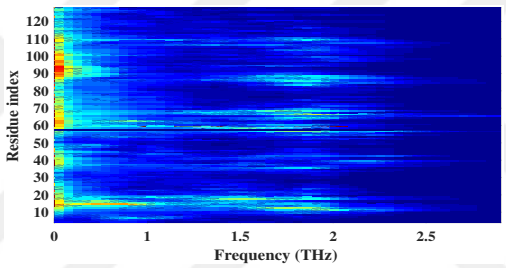
Major-axis length, $|\boldsymbol{r}^{major}{}_i(\omega)|$. Figures 8.5 and 8.12 show the major-axis length both as a function of residue index and frequency, and with averaging over frequency with input residue 57 and 91, respectively. Based on the frequency averaged results, residues 12-18, 56, 66, 68 (87-94, 96, 109) move more than the others and those that exhibit the least amount of movement are the region between 2-5,



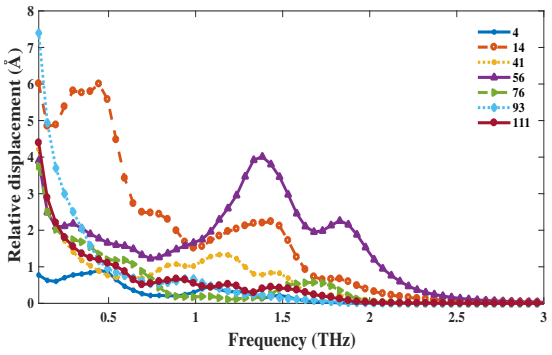
(a) Residue displacements in 3D view



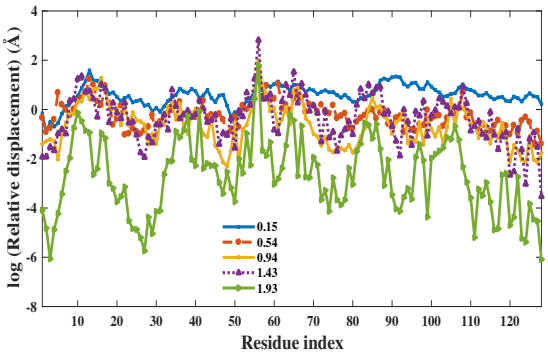
(b) Residue displacements vs residue index



(c) Residue index vs frequency

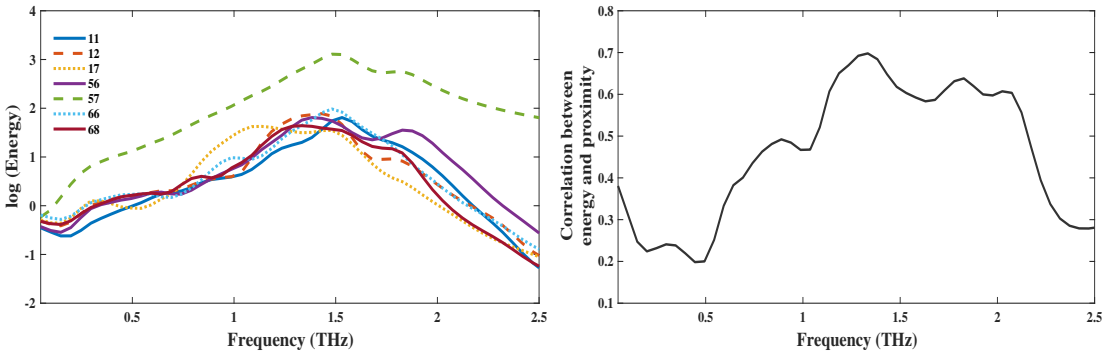


(d) Relative displacements of selected residues



(e) Relative displacements of all residues for selected frequencies

Figure 8.3: Relative displacement magnitudes as a function of residue index and frequency (input residue: 57)

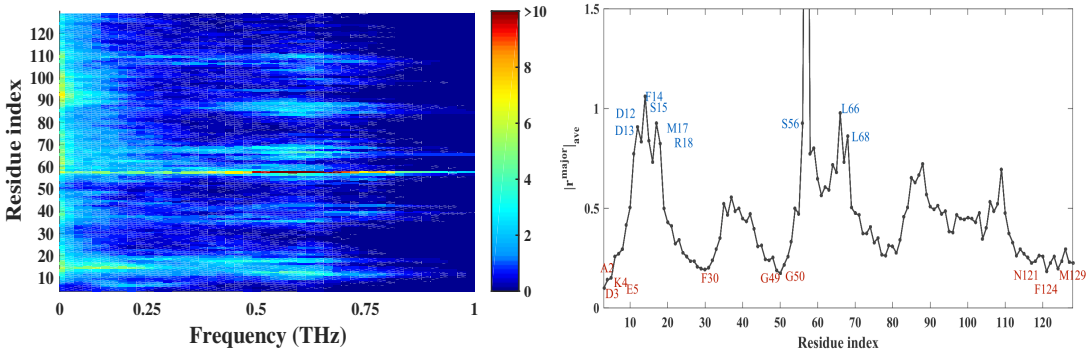


(a) Energy of *critical* residues

(b) Correlation between the energy and the proximity to the input residue as a function of frequency

Figure 8.4: Energy analysis of CheY (input residue: 57)

30, 49, 50, 121, 124, 129 (2, 3, 6, 43, 44, 46, 47, 51, 52, 53) when input residue is 57 (91).



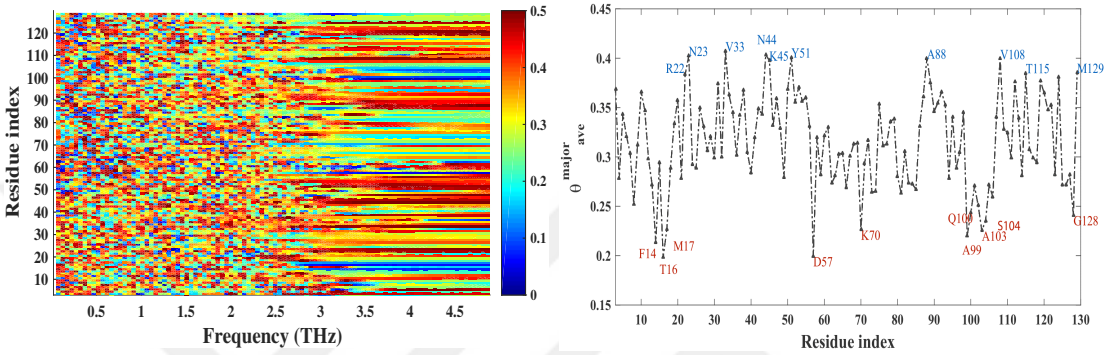
(a) $|\boldsymbol{r}^{major}|$

(b) Frequency-averaged $|\boldsymbol{r}^{major}|_{ave}$. Values above 1.5 Å are not shown.

Figure 8.5: Major-axis length $|\boldsymbol{r}^{major}|$ of CheY (input residue: 57).

Angle between the major-axis and applied force direction, ${\theta^{major}}_i(\omega)$. Figure 8.6 shows that most of the residues tend to be orthogonal to the force for larger frequencies. Frequency averaged θ^{major} values are shown in Figure 8.6(b) and

8.13(b). On average, residues 22, 23, 33, 44, 45, 51, 88, 108, 115, 129 (3, 29, 32, 106, 112, 114, 115, 119, 122, 127) and move orthogonally to the force and residues 14, 16, 17, 57, 70, 99, 100, 103, 104, 128 (4, 10, 17, 18, 41, 82, 88, 89, 91, 109) are aligned with the excitation force direction when input residue is 57 (91).



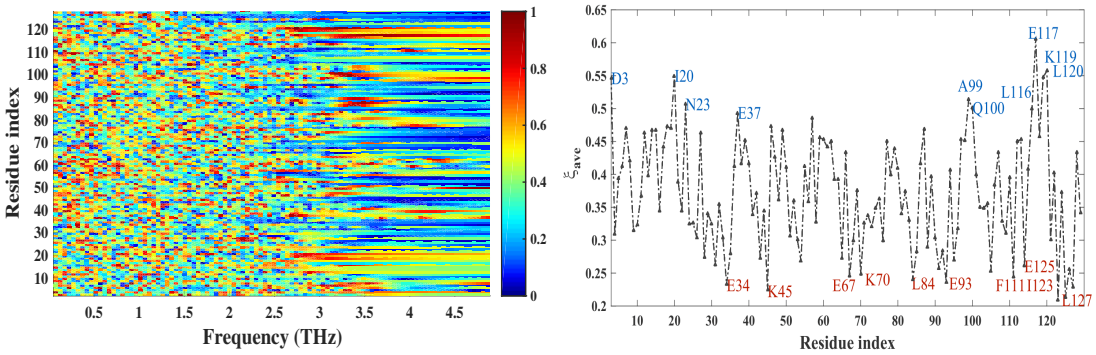
(a) θ^{major} (b) Frequency-averaged $\theta^{major}{}_{ave}$

Figure 8.6: Angle between the major-axis and force, θ^{major}, of CheY. Colorbar labels are to be multiplied with π, and in radians (input residue: 57).

Isotropicity, $\xi_i(\omega)$. Figures 8.7(a) and 8.14(a) show isotropicity as a function of frequency. Over a certain frequency range, most of the residues become anisotropic except residues 115-120 for input residue 57. Figure 4.9(b) shows the frequency-averaged values of isotropicity. Residues 34, 45, 67, 70, 84, 93, 111, 123, 125, 127 (62, 71, 90, 91, 100, 104, 110, 111, 124, 128) are the most anisotropic ones on the average, and 3, 20, 23, 37, 99, 100, 116, 117, 119, 120 (9, 20, 24, 27, 34, 78, 79, 81, 103, 118) are the most isotropic when the force is applied to residue 57 (91).

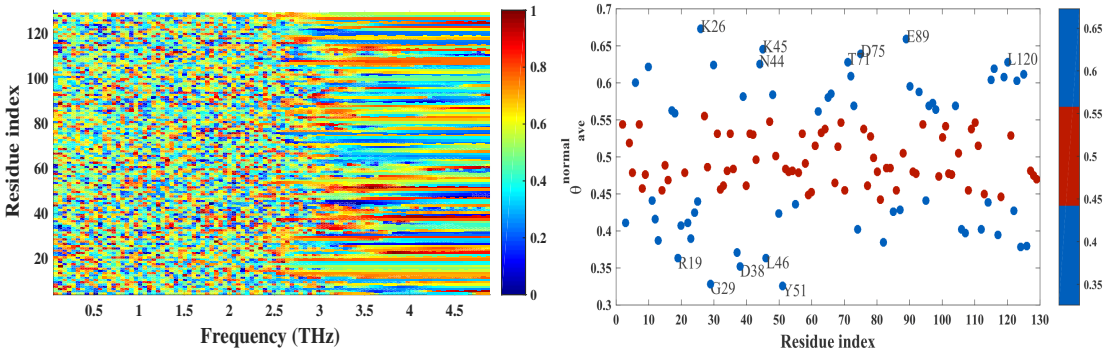
Angle between the plane normal and force, $\theta^{normal}(\omega)$. Figures 8.8(a) and 8.15(a) show the effect of frequency on the coplanarity with two different force excitation sites. When the frequencies are averaged, 19, 26, 29, 44 - 46, 51, 71, 75, 89, 120 (11, 17, 37, 55, 62, 76, 86, 106, 111, 115, 126, 127) are relatively non-coplanar with the force when the input residue is 57 (91).

According to Table 8.1, the single point, static force application seems to model ligand binding better for FBP in comparison to CheY.



(a) ξ (b) Frequency-averaged ξ_{ave}

Figure 8.7: Isotropicity ξ of CheY. Colorbar labels are to be multiplied with π, and in radians (input residue: 57).



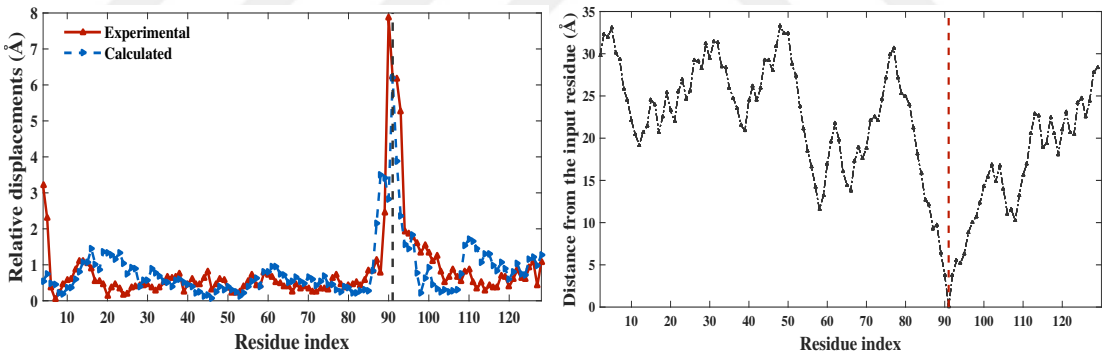
(a) θ^{normal}

(b) Frequency-averaged ${\theta^{normal}}_{ave}$. Red circles indicate the residues whose motions are aligned with the force, and blue circles show the orthogonal motions.

Figure 8.8: Angle between the plane normal and force direction, θ^{normal}, of CheY. Colorbar labels in radians/pi (input residue: 57).

Protein	Unit Force Direction	Force Magnitude	Input Residue	Correlation
FBP	[0.9040 0.3834 0.1894]	4.1859	57	0.9731
CheY	[-0.4060 0.6761 0.6148]	36.8113	57	0.3496
CheY	[0.6485 -0.4982 -0.5705]	11.4416	91	0.6942

Table 8.1: Force parameters, input residue and correlation value between the experimental and calculated displacements for selected proteins.

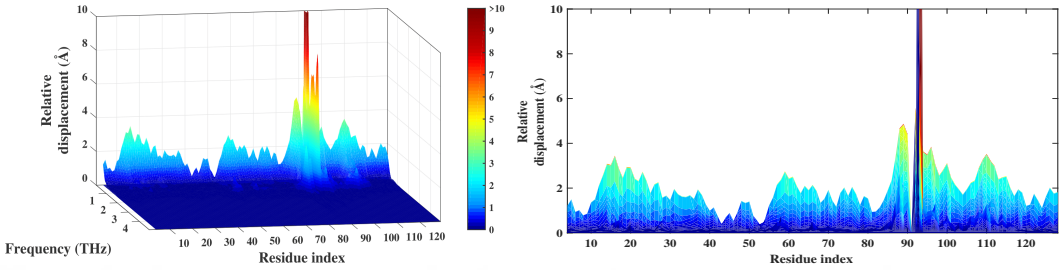


(a) Experimental and calculated displacements between apo and holo forms of CheY. (57)

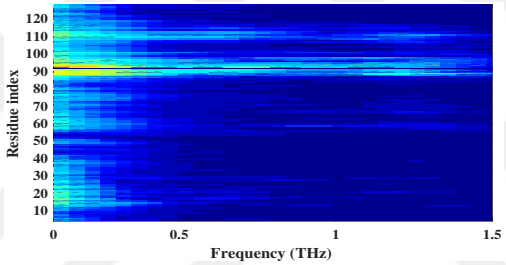
(b) Residue distances to the input residue

Force excitation on residue 91.

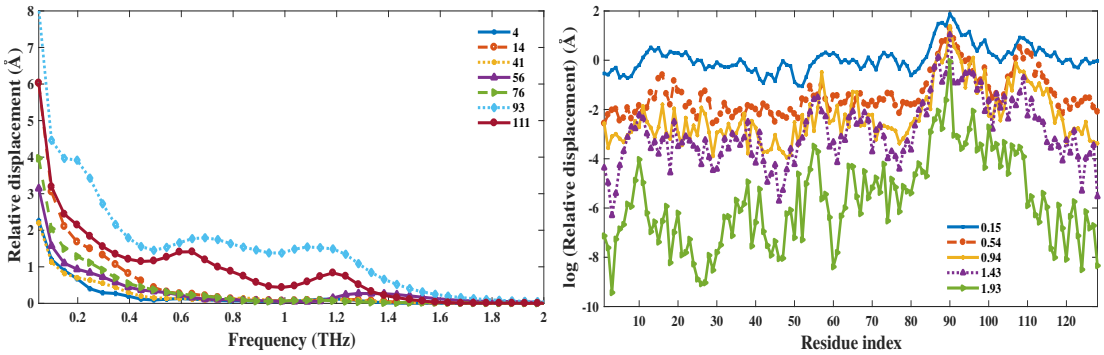
Figure 8.9: Displacement magnitudes of CheY residues in Å. Red curve shows experimental displacements and blue curve is calculated displacements at the best force magnitude and direction at zero frequency. Residues 88, 89, 90, 92, 93, 94 are within 8 Å range of residue 91.



(a) Residue displacements in 3D view (b) Residue displacements vs residue index

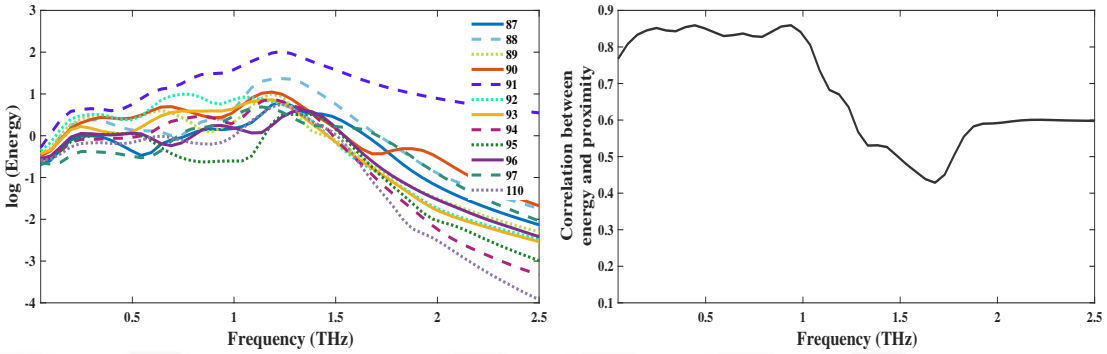


(c) Residue index vs frequency



(d) Relative displacements of selected residues (e) Relative displacements of all residues for selected frequencies

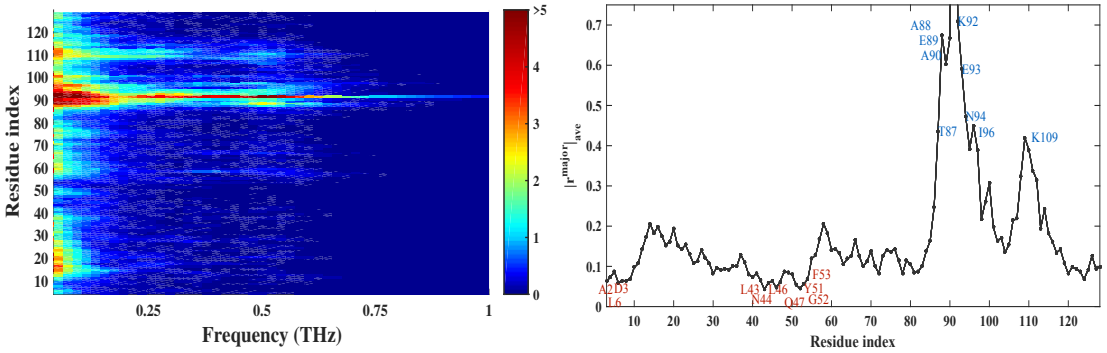
Figure 8.10: Relative displacement magnitudes as a function of residue index and frequency (input residue: 91)



(a) Energy of *critical* residues

(b) Correlation between the energy and the proximity to the input residue as a function of frequency

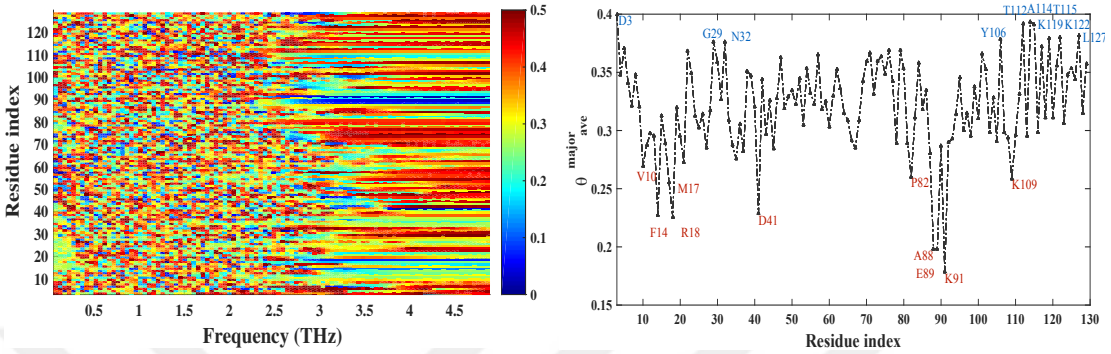
Figure 8.11: Energy analysis of CheY (input residue: 91)



(a) $|\boldsymbol{r}^{major}|$

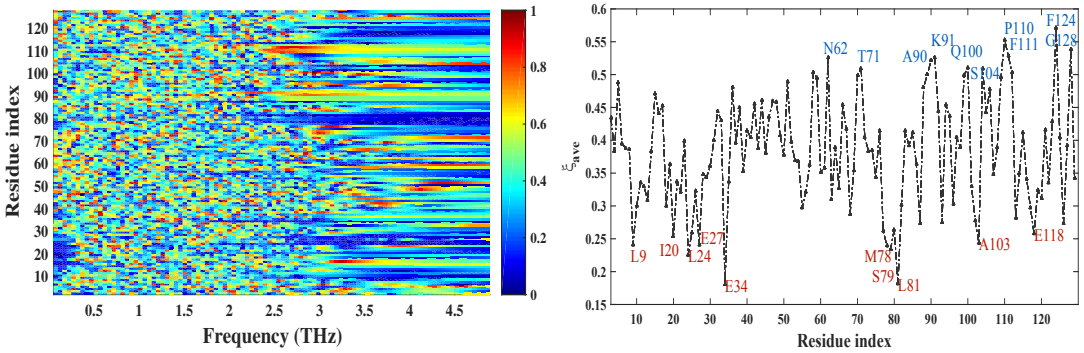
(b) Frequency-averaged $|\boldsymbol{r}^{major}|_{ave}$. The plot is clipped for the input residue for better visualization purposes.

Figure 8.12: Major-axis length $|\boldsymbol{r}^{major}|$ of CheY (input residue: 91).



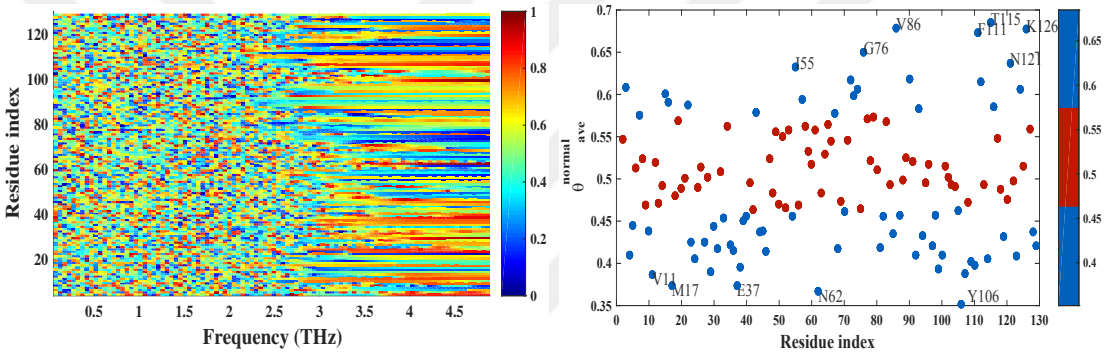
(a) θ^{major} (b) Frequency-averaged ${\theta^{major}}_{ave}$

Figure 8.13: Angle between the major-axis and force, θ^{major}, of CheY. Colorbar labels are to be multiplied with π, and in radians (input residue: 91).



(a) ξ (b) Frequency-averaged ξ_{ave}

Figure 8.14: Isotropicity ξ of CheY. Colorbar labels are to be multiplied with π, and in radians (input residue: 91).



(a) θ^{normal}

(b) Frequency-averaged $\theta^{normal}{}_{ave}$. Red circles indicate the residues whose motions are aligned with the force, and blue circles show the orthogonal motions.

Figure 8.15: Angle between the plane normal and force direction, θ^{normal}, of CheY. Colorbar labels in radians/pi (input residue: 91).

8.2 Supplementary Information for Chapter 5: Allostery in proteins as point-to-point telecommunication in a network: Frequency decomposed signal-to-noise ratio and channel capacity analysis

Amino acid type	**Mass**(amu)	**Volume** ($\times 10^{-3}$ nm^3)
ALA	71	87.2
ARG	156	181.3
ASN	114	117.4
ASP	115	114.6
CYS	103	106.7
GLU	129	141.4
GLN	128	142.4
GLY	57	60.6
HIS	137	152.4
ILE	113	168.9
LEU	113	168.9
LYS	128	174.3
MET	131	163.1
PHE	147	187.9
PRO	97	122.4
SER	87	91.0
THR	101	117.4
TRP	186	228.5
TYR	163	192.1
VAL	99	141.4

Table 8.2: Mass and volume values of the amino acids

Table 8.3 summarizes the critical residues identified in previous work using a wide range of computational methods as well as experimental techniques for the PDZ3

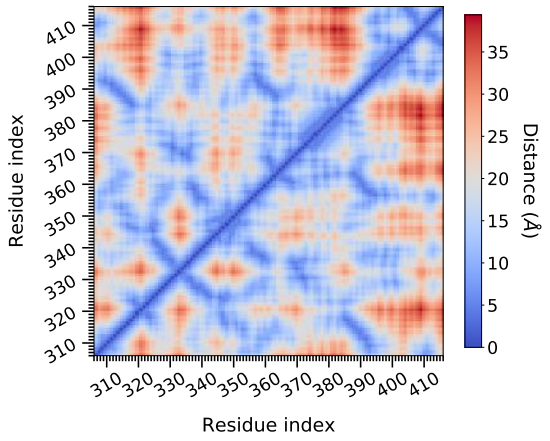
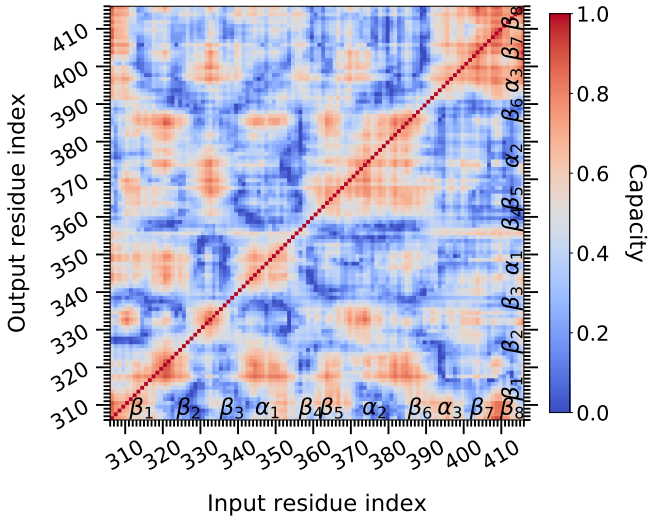


Figure 8.16: Pairwise distances between the residues

domain protein PSD-95. The summaries provided in [116] and [5] were utilized in the preparation of the table.

(a)



(b)

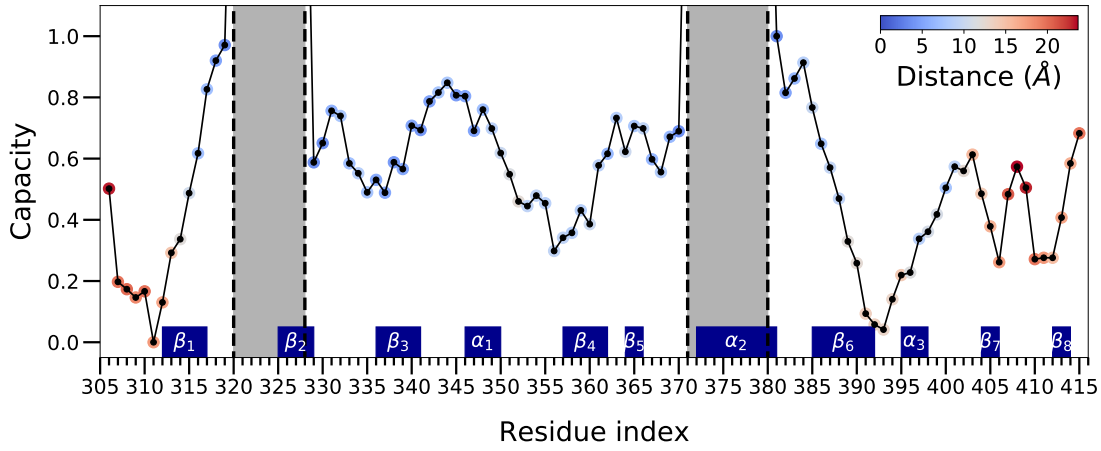
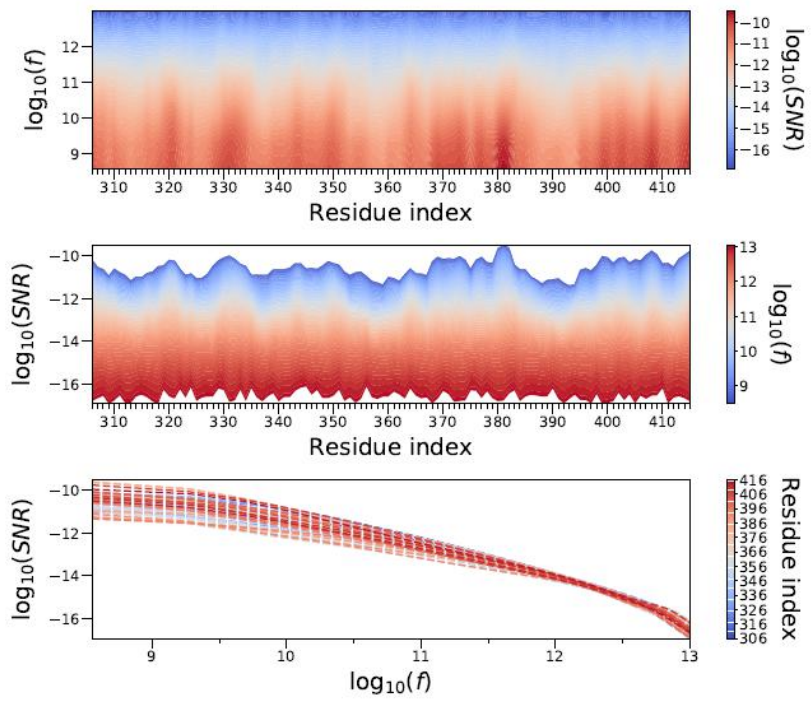


Figure 8.17: (a) Per residue scan and (b) BP excitation with $\boldsymbol{H}^{\mathrm{ANM}}$

(a)


log10(f)
Residue index
log10(SNR)
log10(SNR)
Residue index
log10(f)
log10(SNR)
log10(f)
Residue index

(b)

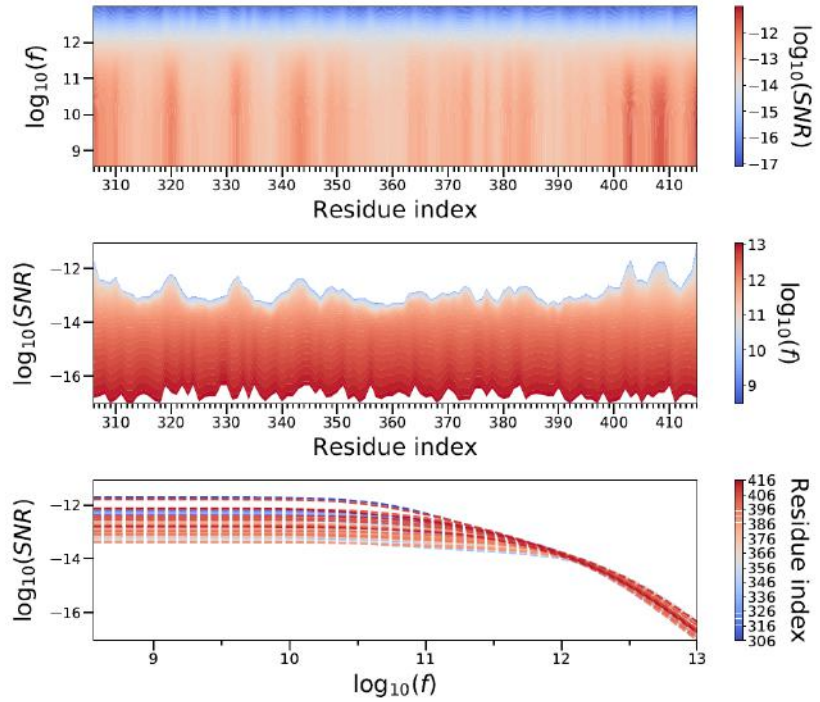


Figure 8.18: PSD profile for the equilibrium fluctuations (a) $\boldsymbol{H}^{\mathrm{MD}}$ (b) $\boldsymbol{H}^{\mathrm{ANM}}$

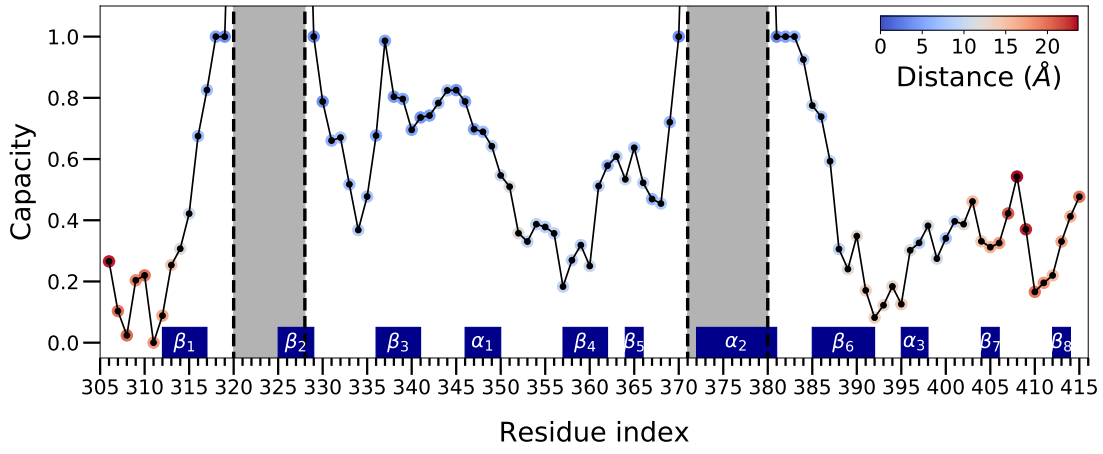
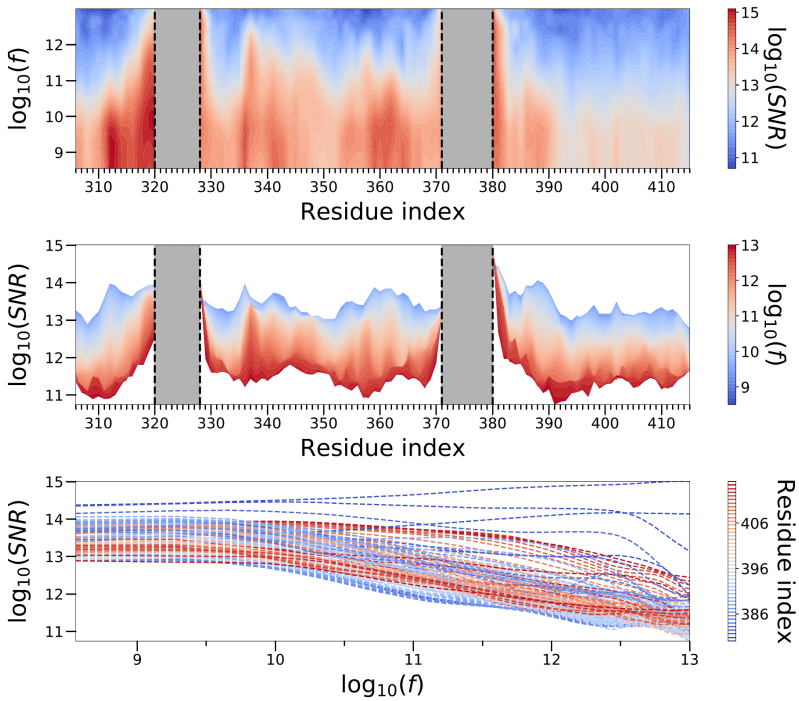


Figure 8.19: Normalized capacity values in BP excitation mode using $\boldsymbol{H}^{\mathrm{MD}}$. The colorbar shows the minimum distance between a residue and one of the BP residues.

(a)


log10(f)
log10(SNR)
Residue index
log10(SNR)
log10(f)
Residue index
log10(SNR)
log10(f)
Residue index

(b)

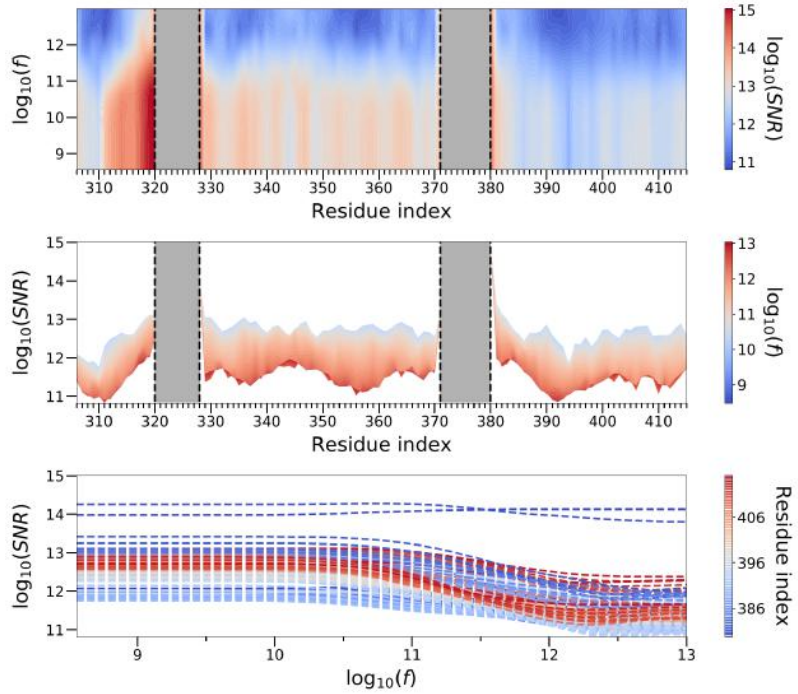


Figure 8.20: Binding pocket excitation with (a) $\boldsymbol{H}^{\mathrm{MD}}$ (b) $\boldsymbol{H}^{\mathrm{ANM}}$.

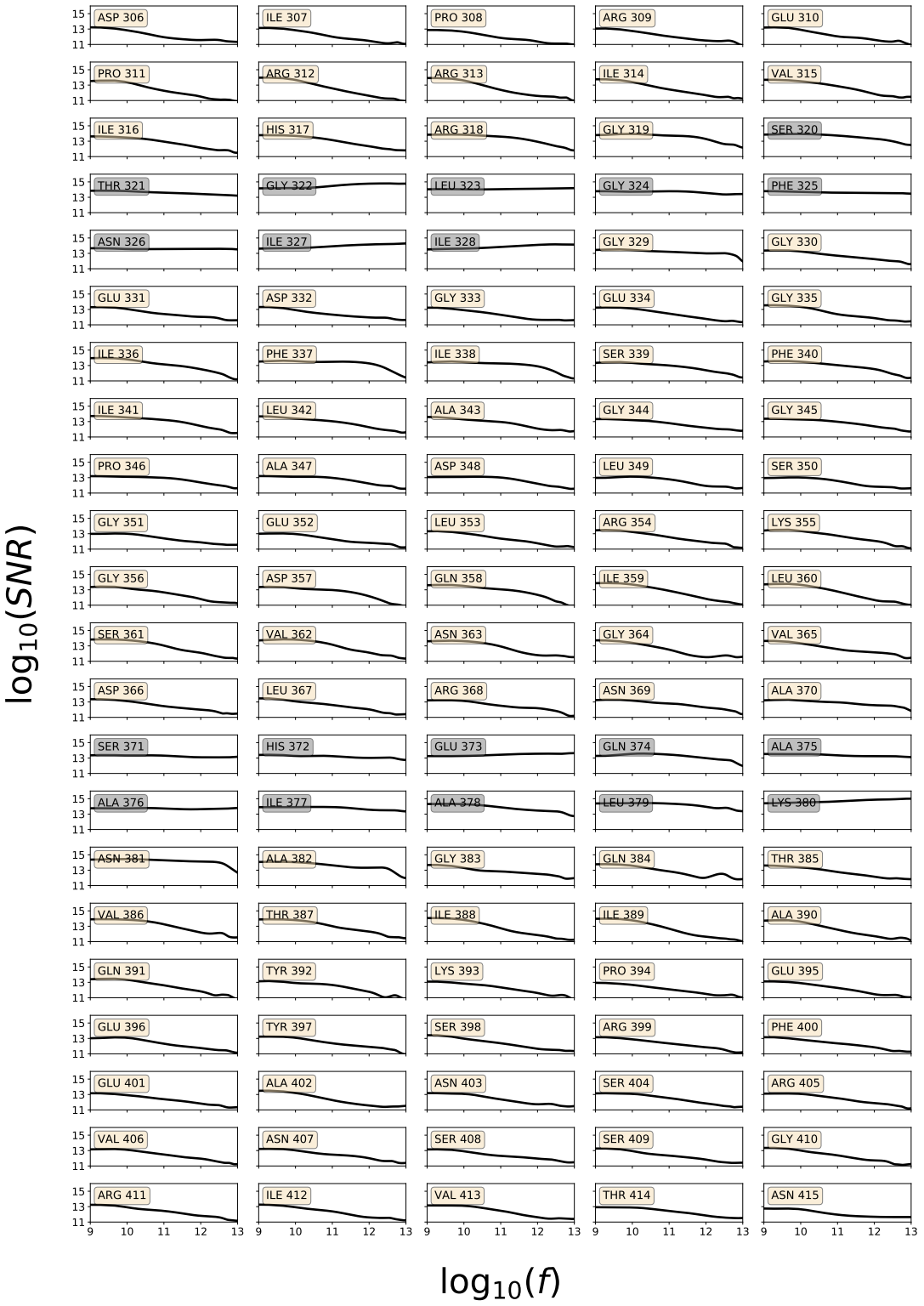


Figure 8.21: Individual frequency response of the residues in BP excitation with $\boldsymbol{H}^{\mathrm{MD}}$.

Method	**Critical residues**
Perturbation response scanning [5]	314, 316, 326, 327, 328, 329, 330, 335, 336, 337, 338, 339, 340, 341, 345, 346, 347, 353, 354, 355, 356, 358, 359, 361, 362, 367, 370, 372, 375, 379, 386, 387, 388, 389, 390
Experimental [6]	325, 328, 329, 340, 341, 362, 372, 376, 380, 386, 390
Statistical coupling analysis. 1 [7]	322, 325, 329, 340, 341, 362, 372, 376, 380, 386
Statistical coupling analysis. 2 [8]	323, 324, 325, 327, 328, 329, 330, 336, 338, 341, 347, 353, 359, 362, 367, 372, 375, 376, 379, 388
Anistotropic thermal diffusion [9]	325, 327, 341, 347, 353, 372
Structural perturbation method [10]	310, 318, 319, 320, 323, 327, 329, 331, 332, 333, 334, 372, 376, 380, 384, 400
Rotamerically induced perturbation [11]	316, 318, 323, 325, 336, 346, 347, 349, 353, 357, 359, 362, 367, 375, 378, 379, 386, 390
MD [12]	309, 318, 322, 323, 324, 326, 327, 328, 330, 331, 332, 334, 337, 348, 352, 354, 355, 357, 373, 380, 391, 395, 399
Deep coupling scan [13]	372, 375, 376, 379
Thermodynamic double mutant cycle [14]	314, 316, 323, 325, 327, 328, 331, 340, 341, 347, 353, 362, 372, 375, 376, 377, 378, 380, 382, 386
Conservation mutation correlation analysis [15]	311, 316, 326, 327, 330, 336, 337, 338, 339, 341, 342, 343, 344, 347, 348, 351, 356, 359, 360, 362, 364, 366, 369, 376, 378, 381, 386
Rigid-residue scan [16]	338, 343, 347, 397
Monte Carlo path [17]	325, 327, 338, 353, 372

Table 8.3: Critical residues identified by various methods

8.3 Supplementary Information for Chapter 6: Interpretable embeddings from molecular simulations using Gaussian mixture variational autoencoders

X′ ∈ $\mathbb{R}^n$ denotes the reconstructions. The sampling operation in the reconstructions (shown in the decoding part of Figure 6.3), corresponds to taking the means of the Gaussians for simplicity.

8.3.1 One dimensional 4-well potential

The trajectory data is obtained as suggested in [155], and using the code provided in [163]. 100 × 100 transition probability matrix is obtained among the equally-spaced 100 bins in the interval [-1, 1] as follows

$$P_{ij} = \begin{cases} C_i \exp\left(-(V_i - V_j)\right), & \text{if } |i-j| \leq 1 \\ 0, & \text{otherwise ,} \end{cases} \tag{8.1}$$

where V_i and V_j are the potential energies at the centers of bins i and j, which are defined according to the potential of the form: $V(X) = 2(X^8 + 0.8e^{-80X^2} + 0.2e^{-80(X-0.5)^2} + 0.5e^{-40(X+0.5)^2})$, and C_i is the normalization factor. The system is initialized randomly, and propagated according to P_{ij} 5×10^6 steps in time.

Figure 8.22 shows the reconstructions in a scatter plot. The $X = X'$ line shows the lossless reconstructions.

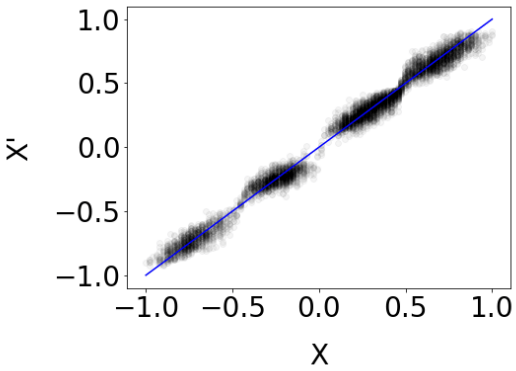


Figure 8.22: X vs X′ for one-dimensional 4-well potential. Reconstructions are obtained via the GMVAE.

8.3.2 Müller-Brown potential

The trajectory data is obtained as suggested in [131], and using the code provided in [164]. Two dimensional potential energy is defined as:

$$V(X_0, X_1) = \sum_{j=0}^{3} A_j \exp[a_j(X_0 - x_{0,j})^2 + b_j(X_0 - x_{0,j})(X_1 - x_{1,j}) + c_j(X_1 - x_{1,j})^2] \ , \tag{8.2}$$

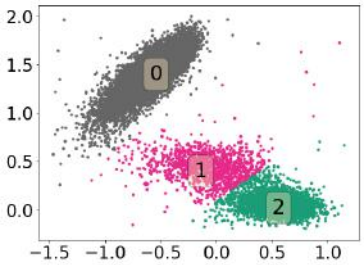
where $x = (X_0, X_1)$ is the two-dimensional coordinate, and A, a, b, c, x_0 and y_0 are the standard parameters [152] such that $A = (-200, -100, -170, 15)$, $a = (-1, -1, 6.5, 0.7)$, $b = (0, 0, 11, 0.6)$, $c = (-10, -10, -6, 5, -0.7)$, $x_0 = (1, 0, -0.5, -1)$, $x_1 = (0, 0.5, 1.5, 1)$. The trajectory data is generated using 30 trajectories of 10000 steps simulated with Brownian dynamics:

$$\frac{dx}{dt} = -\frac{\Delta V(x)}{kT} + \sqrt{2D}R(t) \ , \tag{8.3}$$

where $kT = 1.5 \times 10^4$ joules, and $D = 10^{-2}$ meters-squared per second, and $R(t)$ is a delta-correlated Gaussian process with zero mean.

The true labels are defined as shown in Figure 8.23(a). Figures 8.23(b) and 8.23(c) show the reconstructions.

(a) True labels



(b) X0 vs X0′ (c) X1 vs X1′

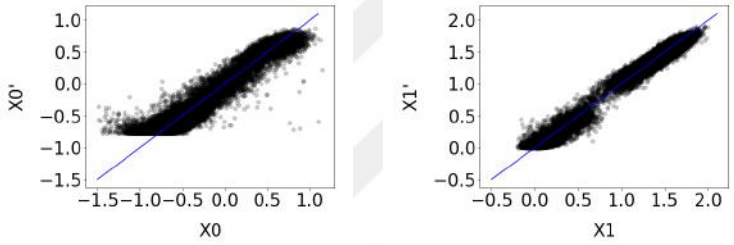


Figure 8.23: True label definitions and X vs X′ for Müller-Brown potential. Reconstructions are obtained via the GMVAE. (a) True labels. (b) Reconstructions in the first dimension. (c) Reconstructions in the second dimension.

Figure 8.24 further demonstrates the ability of the GMVAE to learn a nonlinear manifold that separates the three distinct free-energy basins, compared with time-lagged independent component analysis (TICA), which can only find a linear separatrix for the basins. Figure 8.24, showing the projections obtained with the GMVAE and TICA, was constructed following [131], with the colors indicating values of the latent variable while the gray dots correspond to trajectory data.

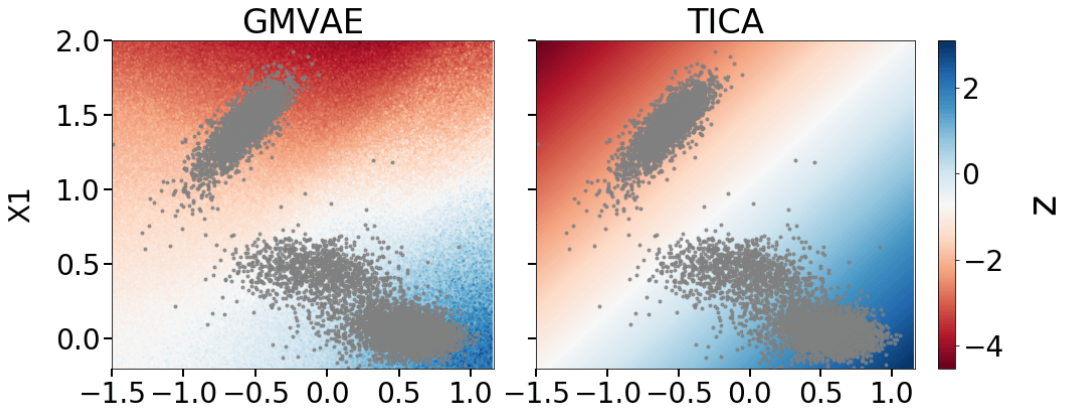
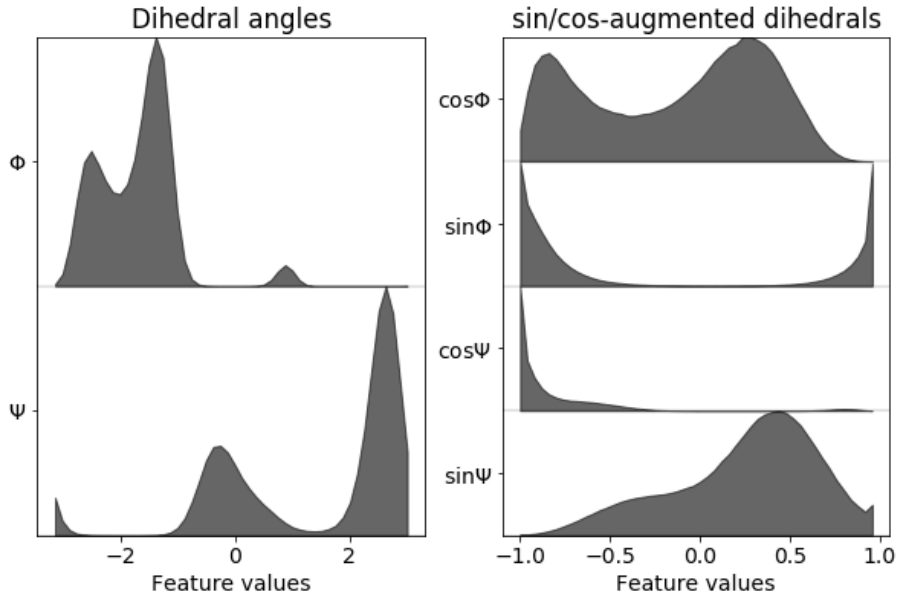


Figure 8.24: Projections via the GMVAE and TICA. The GMVAE learns the non-linear dividing surface in the low-dimensional space.

8.3.3 Alanine dipeptide

As the input features, dihedral angles and pairwise distances for heavy atoms that are provided in [156] for three simulations of length 250 ns each are used. Dihedral angles are transformed to their sin / cos representaions, and the pairwise distances whose variance are low are removed from the feature set (using kurtosis function from scipy.stats library [165], with threshold value of 0.03, as shown in Figure 8.25).

(a) Dihedral angles



(b) Pairwise distances

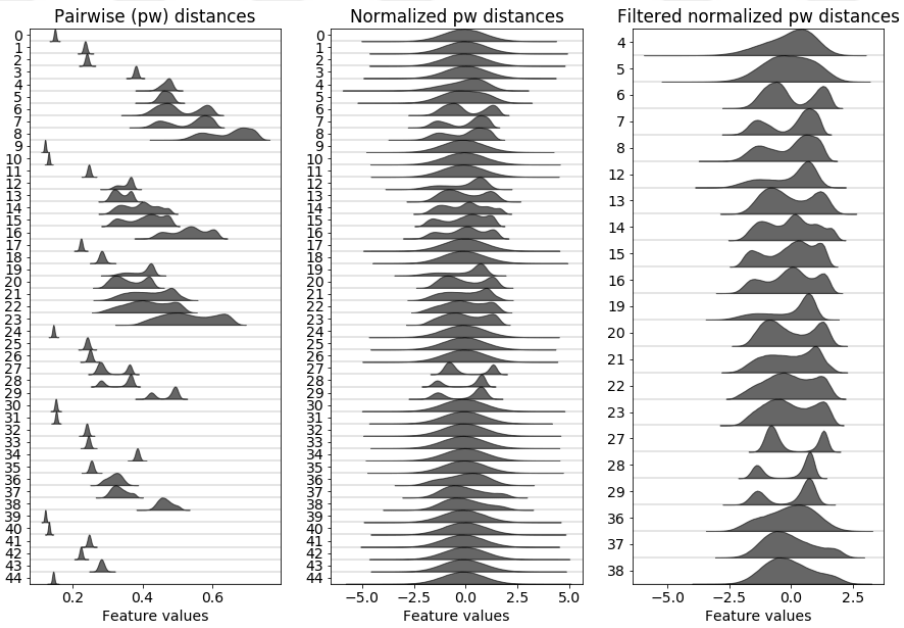


Figure 8.25: Processing of the features: (a) the dihedral angles along the backbone, and (b) the pairwise distance between heavy atoms.

We applied TICA to the set of pairwise distances only, followed by a kinetic

coarse-graining with the PCCA$^+$ method into 4 metastable states. Figure 8.26(a) presents the resulting clusters plotted on the Ramachandran plot. Figures 8.26(b) and 8.26(c) show the histograms of these metastable states, and the GMVAE clusters, respectively.

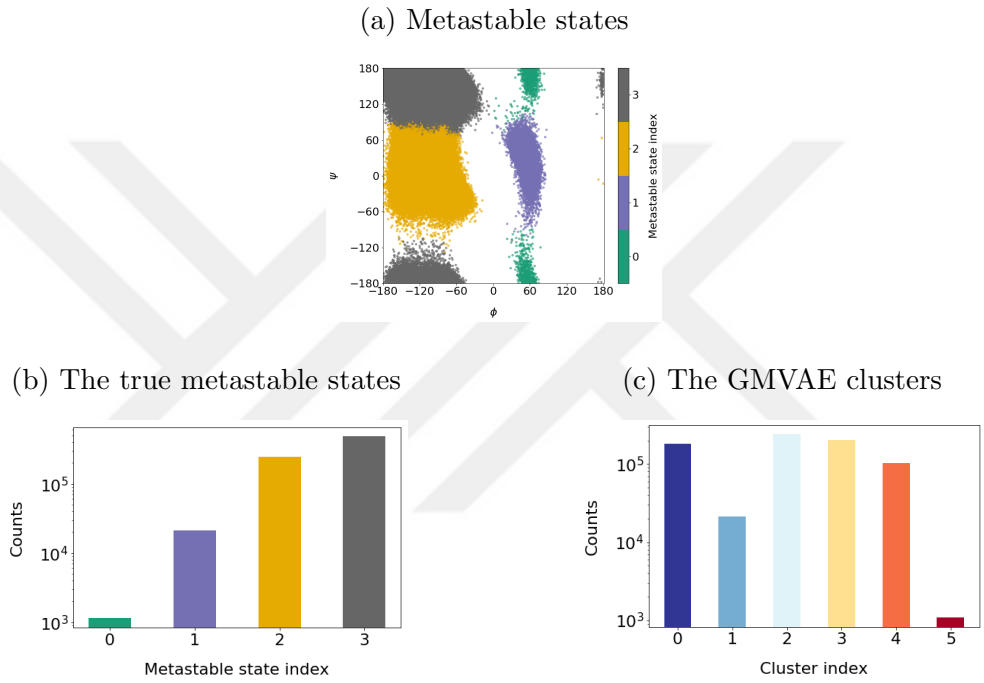


Figure 8.26: (a) Metastable states from TICA and PCCA$^+$ on the Ramachandran plot. The histograms for the (b) true metastable states, (c) GMVAE clusters.

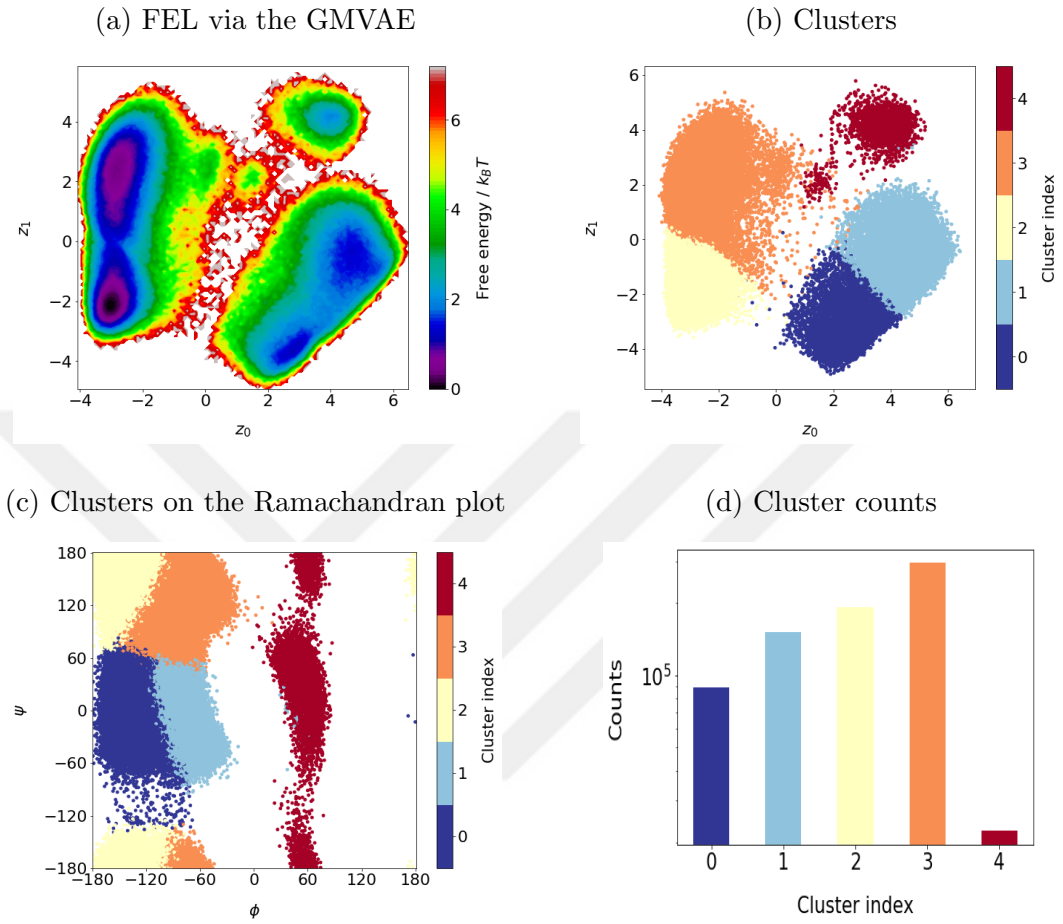


Figure 8.27: (a) FEL obtained for the alanine dipeptide by the GMVAE in a separate fully-converged training. The GMVAE clusters on the (b) GMVAE landscape, (c) Ramachandran plot. (d) Cluster counts.

(a) Implied timescales (b) Chapman-Kolmogorov test

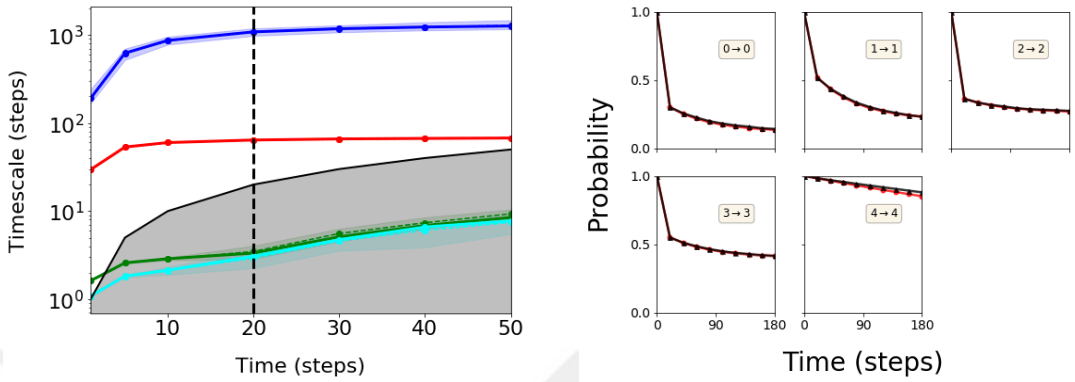


Figure 8.28: Markovianity check of the MSM built for alanine dipeptide via the GMVAE. (a) Implied timescales. (b) Chapman-Kolmogorov test (at lag=20 steps).

(a) FEL via the VAE (b) The true metastable state partitions

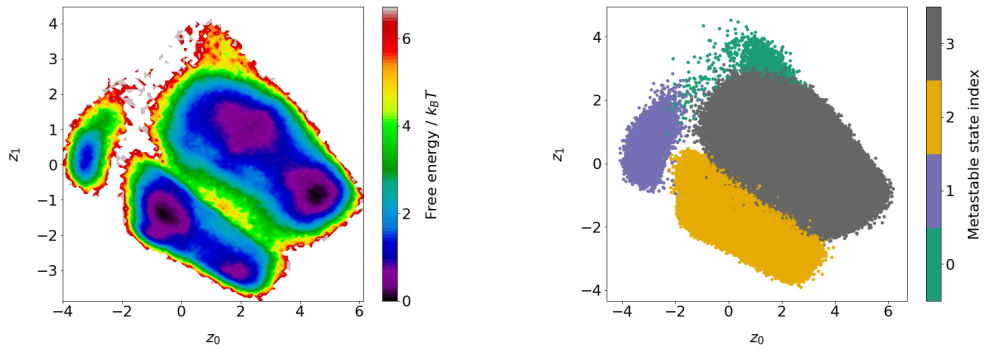


Figure 8.29: VAE results for alanine dipeptide. (a) The FEL obtained by the VAE, (b) the true metastable state partitions on this landscape.

8.3.4 $AAQAA_3$ peptide - I

8.3.4.1 Coarse-grained peptide model

We employ a simple physics-based peptide model that was previously used to investigate structural-kinetic relationships in helix-coil transitions [159, 166]. The model employs three attractive interactions, following standard Gō-type models [167,

168, 169, 170]: *(i)* a native contact (nc) attraction, U_{nc}, employed between pairs of $C_\alpha$atoms which lie within a certain distance in the native structure, i.e., the α-helix, of the peptide, *(ii)* a desolvation barrier (db) interaction, U_{db}, also employed between native contacts, and *(iii)* a hydrophobic (hp) attraction, U_{hp}, employed between all pairs of C_β atoms of the amino acid side chains. We employed the same functional forms as in many previous studies [170], with a tunable prefactor, ϵ_i, for each of the interactions. The model considered here employed the prefactors $\epsilon_{\mathrm{nc}} = 12.5$, $\epsilon_{\mathrm{db}} = 0.4\epsilon_{\mathrm{nc}}$, and $\epsilon_{\mathrm{hp}} = 0.2\epsilon_{\mathrm{nc}}$, while performing simulations at a temperature of 280 K. In addition to these simple coarse-grained interactions, a standard AA force field, AMBER99sb [103], is also partially incorporated to model both the steric interactions between all non-hydrogen atoms and also the specific local conformational preferences along the chain.

Molecular dynamics simulations of $AAQAA_3$ were performed with the Gromacs 4.5.3 simulation suite [102] in the constant NVT ensemble, while employing the stochastic dynamics algorithm with a friction coefficient $\gamma = (2.0\ \mathcal{T}^{\mathrm{S}})^{-1}$ and a time step of $1 \times 10^{-3}\ \mathcal{T}^{\mathrm{S}}$. For each model, 100 independent simulations were performed with starting conformations varying from full helix to full coil. Each simulation was performed for $100,000\ \mathcal{T}^{\mathrm{S}}$, recording the system every $0.5\ \mathcal{T}^{\mathrm{S}}$. The CG unit of time, $\mathcal{T}^{\mathrm{S}}$, can be determined from the fundamental units of length, mass, and energy of the simulation model, but does not provide any meaningful description of the dynamical processes generated by the model. In this case, $\mathcal{T}^{\mathrm{S}} = 1$ ps.

8.3.4.2 GMVAE landscape and the cluster assignments

Since the GMVAE method is a probabilistic clustering method, each data point has a probability of assignment to each of the k clusters. For a data point d_i, the probability of assignment to each of the clusters has probability values $p_{d_i,0}, p_{d_i,1}, \ldots p_{d_i,k-1}$ for k number of clusters. In the ideal case, all of the probability values except the true cluster is equal to 0, and the true cluster has a value of 1. Figure 8.30 separately shows the histogram of probability distributions of all of the data points for a cluster. For instance, for cluster 0, the probability distributions are accumulated

at probability values 0 and 1. In other words, with high certainty, cluster 0 is differentiated from the others. None of the data points is assigned to clusters $7-8-9$. Note that although the network is initially trained for 10 clusters, it is not possible to separate more than 7 clusters under the specified loss function. This suggests a way to find the inherent number of clusters, i.e., metastable states, provided that k is chosen larger than that true value.

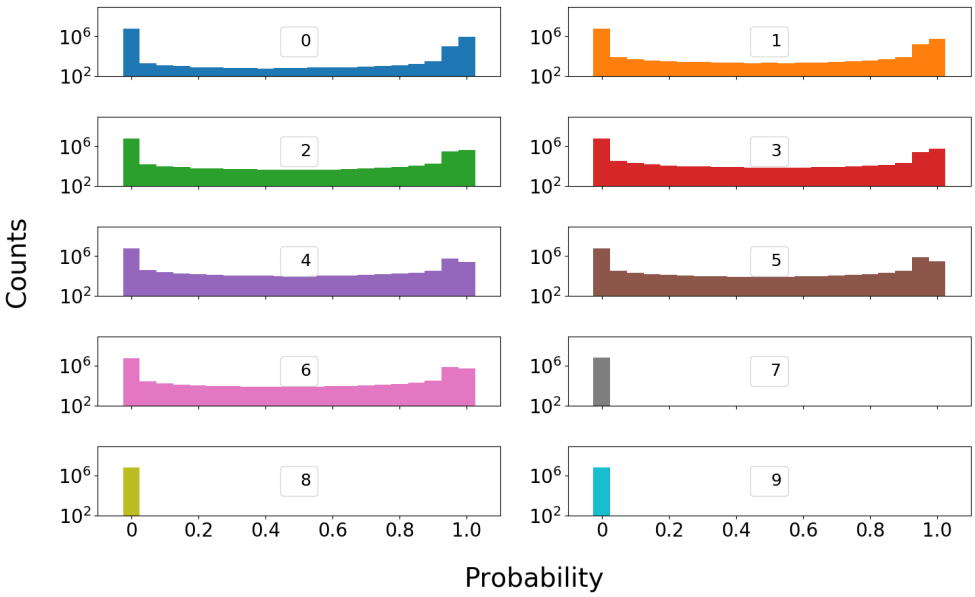


Figure 8.30: The population distribution as a function of probability of belonging to each of the clusters after the training. None of the data points is assigned to clusters $7-8-9$.

Cluster ID's are obtained after a thresholding step as explained in Section 6.2.3.1. Figure 8.31 shows the cluster populations. Cluster -1 indicates the datapoints that are not assigned to any of the clusters.

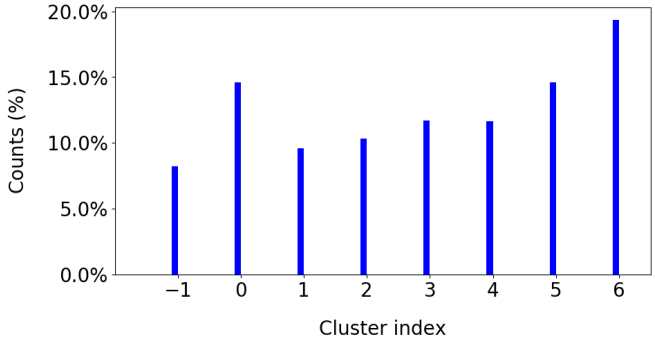


Figure 8.31: Cluster populations

Figure 8.32 shows the inter-cluster $\langle f_h \rangle$ distributions.

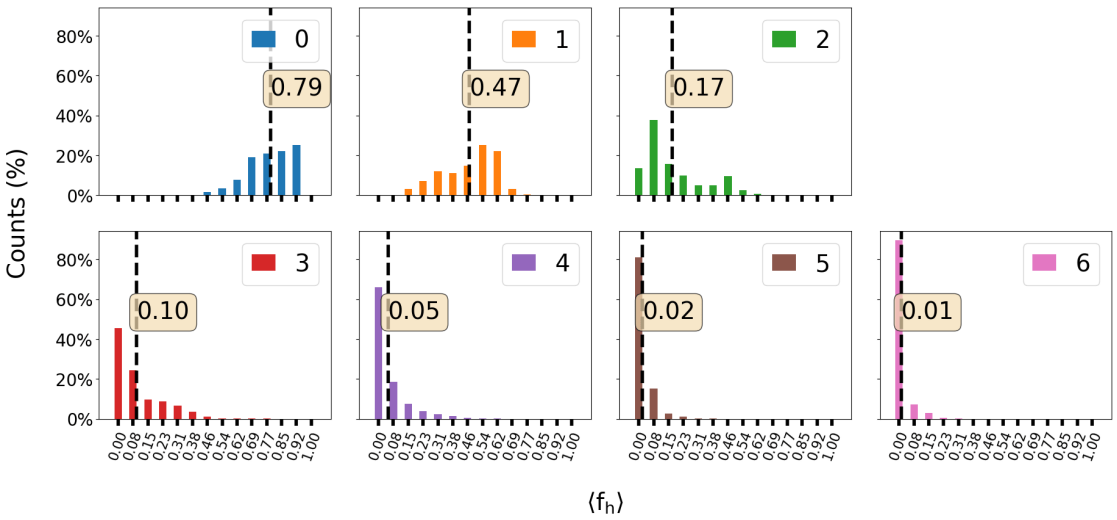
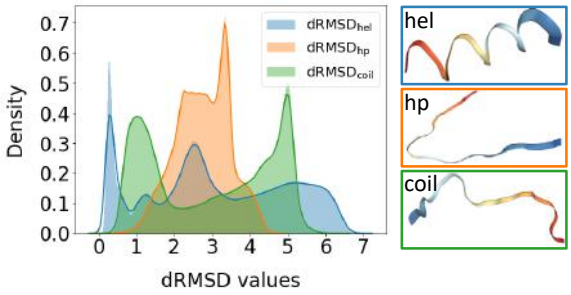


Figure 8.32: Intra-cluster distributions of average helical fraction, $\langle f_h \rangle$, ($AAQAA_3$ - I). The dashed lines indicate the average values, which are also written in the text boxes.

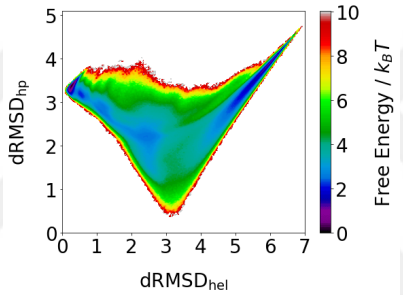
To further characterize the clustering of secondary structures, we separately calculated dRMSDs with respect to three reference structures: helix (hel), hairpin-like (hp), and extended (coil). Figure 8.33(a) presents both the reference structures (right) and corresponding dRMSD distributions (left). The first and the second small peaks in the $dRMSD_{hel}$ distribution represent helical conformations, while the

peak corresponding to $\text{dRMSD}_{\text{hel}}$ values between $2-3.5$ hints at the presence of the hairpin-like structures. Note that there is an offset in dRMSD_{hp} values due to (i) the scarcity of the well-defined hairpins in the trajectory data, and (ii) the subjectivity involved in choosing the reference structures. By plotting the two-dimensional free-energy surface along $\text{dRMSD}_{\text{hel}}$ and dRMSD_{hp}, shown in Figure 8.33(b), the distinct secondary structures can be separated. The conformations with $\text{dRMSD}_{\text{hel}}$ values below 1.8 are helical, whereas the minimum in the upper right with $\text{dRMSD}_{\text{hel}}$ greater than 4 is comprised of extended structures. The energy minimum in the middle (enclosed in the region with $\text{dRMSD}_{\text{hel}}$ values between 1.8 and 3, and dRMSD_{hp} values between 0 and 3) contains hairpin-like structures. Figure 8.33(c) presents inter-cluster free-energy surfaces along $\text{dRMSD}_{\text{hel}}$ and dRMSD_{hp}, generated by considering only conformations within a single cluster.

(a) Histogram plots of dRMSD values



(b) Scatter plot of dRMSD$_{hel}$ vs. dRMSD$_{hp}$



(c) Sampled regions of dRMSD$_{hel}$ vs. dRMSD$_{hp}$ for each cluster

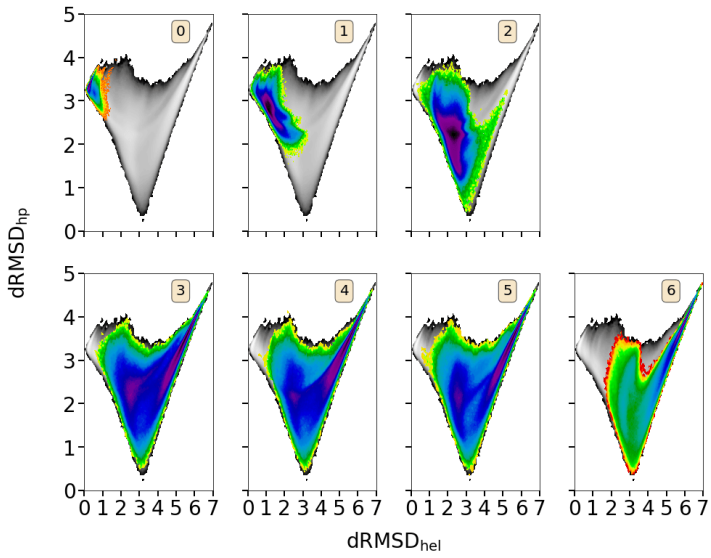


Figure 8.33: dRMSD analysis. (a) The density histograms of dRMSD values for helix, hairpin, and coil structures (with the visualized reference structures). (b) Scatter plot of dRMSD$_{hel}$ vs. dRMSD$_{hp}$ with densities. (c) Sampled regions of dRMSD$_{hel}$ vs. dRMSD$_{hp}$ in each of the cluster. The same colormap is used in (b) and (c).

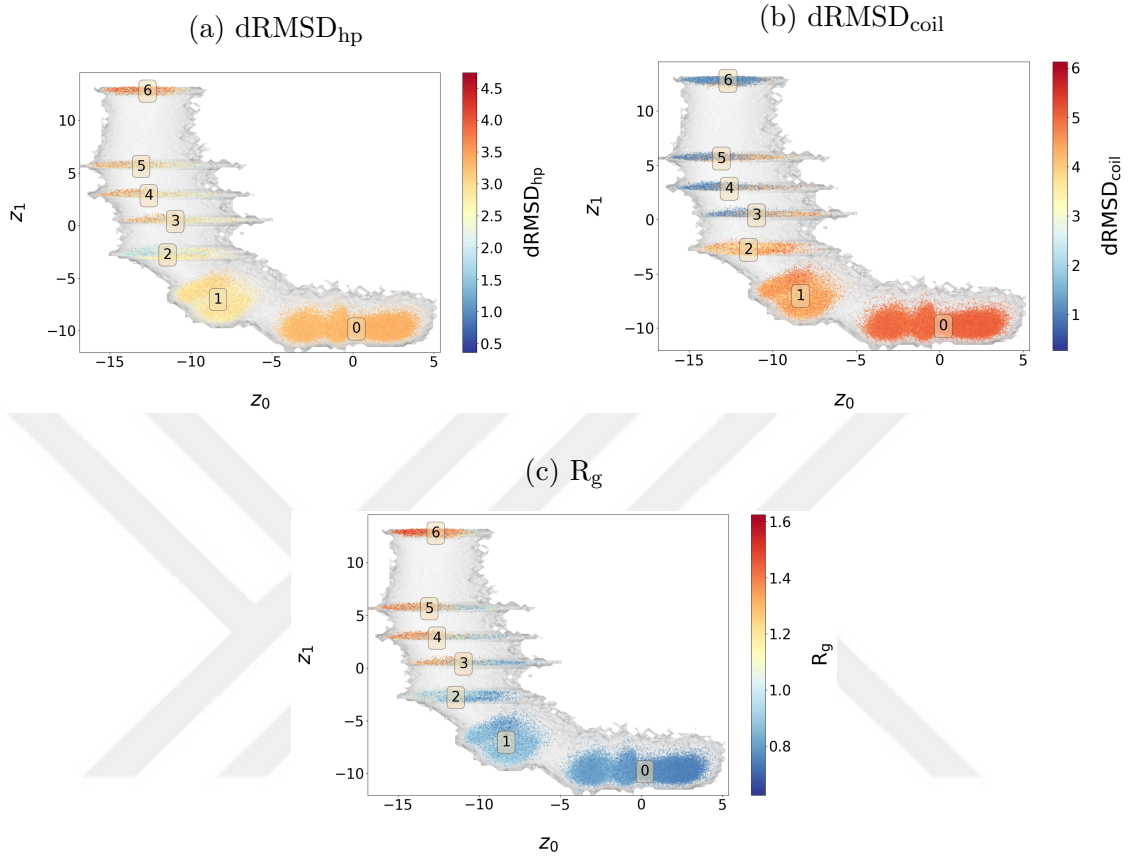


Figure 8.34: Projections colored according to (a) dRMSD$_{hel}$, (b) dRMSD$_{hp}$, (c) dRMSD$_{coil}$, and (d) R$_g$

In addition to $\langle f_h \rangle$ and dRMSD$_{hel}$, the *radius of gyration* R$_g$ distribution is also analyzed. R$_g$ measures the mass-weighted deviations from center of mass, and gives an idea on the overall spread and compactness of the molecule, and is calculated as

$$R_g = \sqrt{\frac{\sum_i m_i ||\mathbf{r}_i - \mathbf{r}_c||^2}{\sum_i m_i}} \,, \tag{8.4}$$

where m_i is the mass of atom i, $\mathbf{r}_i$ the coordinates of atom i, and $\mathbf{r}_c$ is the coordinates of the center of mass. Figures 6.13(b), 8.34(a), 8.34(b), and 8.34(c) show the heat map of dRMSD$_{hel}$, dRMSD$_{hp}$, dRMSD$_{coil}$, and R$_g$ on the FEL obtained via the GMVAE, respectively.

As a final characterization of the clustering, we constructed an MSM directly from the discretized trajectories of GMVAE cluster indices. Although threshold-

ing was applied in the results presented here (practically similar to coring methods for constructing kinetic models [162]), we found that this procedure had negligible effect on the accuracy of the resulting MSM. Figure 8.35(a) presents implied timescale test. The kinetic model can resolve two of longest characteristic processes and demonstrates reasonable convergence, although there is a small increase in the longest timescale with increasing lag time. This subtle discrepancy already indicates that there may be some issues with the accuracy of the kinetic model. An MSM is constructed at lag time 700 to balance between the convergence of the timescales and the resolution of shorter timescale processes. Figure 8.35(b) presents the CK test from this model. There are significant errors in the description of probability decay from each of the metastable states (i.e., clusters), especially states 0 and 1. First, we note that coarse-grained MSMs (i.e., MSMs built on a small number of metastable states) are often not expected to be quantitatively accurate due to difficulties in accurately defining the dividing surfaces between states [143]. However, we anticipate that it should be possible to make a more accurate coarse-grained MSM for this particular simulation trajectory. The discrepancies in the model can then originate from two coupled problems: (i) the GMVAE latent space definition places structures close together that are kinetically distinct (i.e., there are hidden barriers) or (ii) the GMVAE clustering fails to identify/separate distinct metastable states. The FEL within the latent space (Figure 6.11(a)) contains clearly separated basins that are not identified as unique clusters by the GMVAE. In particular, within clusters 0 and 1, there seems to be 2 and 3 separate states, respectively. Figure 8.32 shows that cluster 0 (1) contains structures with a range of helicities ranging from 0.46-1.0 (0.15-0.69). According to the conventional picture of the helix-coil transition, the overarching kinetics can be described by two timescales: (i) the rate at which a single helical segment is formed and (ii) the elongation rate of helical segments along the chain. By grouping together conformations with a single helical segment and several helical segments, the GMVAE has convoluted these two timescales, resulting in non-Markovianity in the kinetics described on these clusters. To further clarify the source of these errors, we constructed an MSM in the conventional way, directly from

the latent space distribution. More specifically, we applied k-means algorithm with 1000 cluster centers, and then applied PCCA$^+$ [146]. In order to enable comparison, we continued with the previous number of metastable states (7). The CK test for the resulting model with lag time $\tau = 700$ (obtained from the implied timescale test, Figure 8.37(b)) is presented in Figure 8.37(c), and demonstrates slightly improved accuracy.

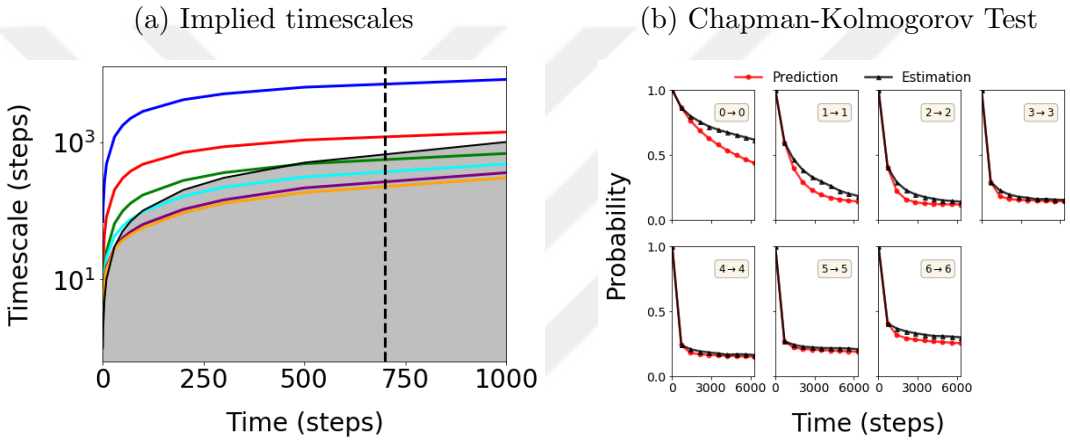


Figure 8.35: Markovianity check of the MSM built for the $AAQAA_3$ peptide - I via using the cluster labels from the GMVAE. (a) Implied timescales. (b) Chapman-Kolmogorov test (at lag=700 steps)

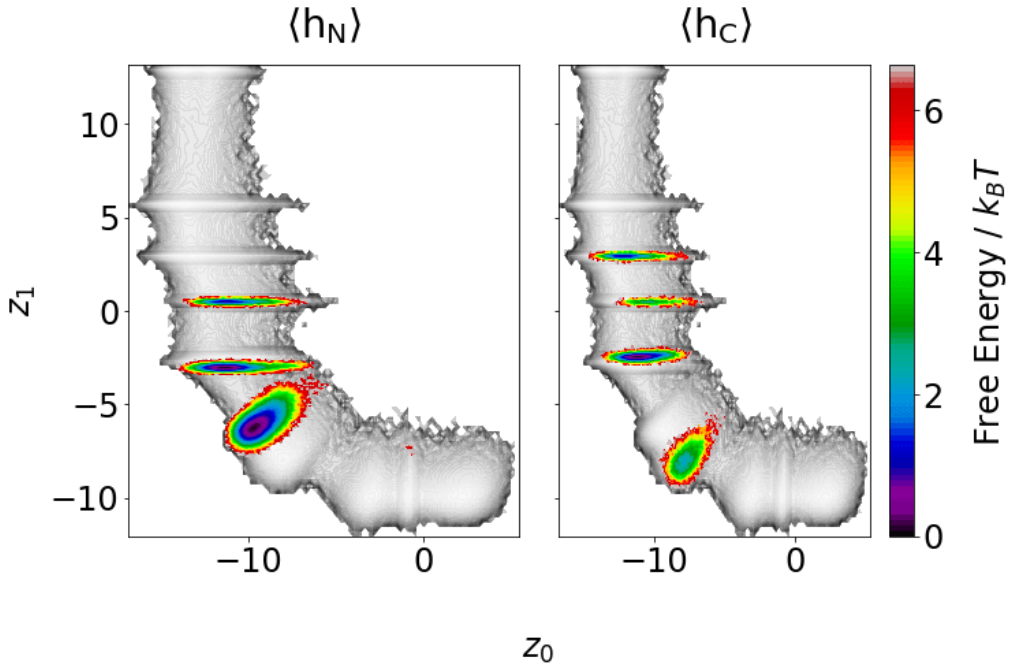


Figure 8.36: Distributions for $\langle h_N \rangle \geq 0.8$ (on the left), $\langle h_N \rangle \leq -0.8$ (on the right).

8.3.4.3 GMVAE landscape only (without using the cluster assignments)

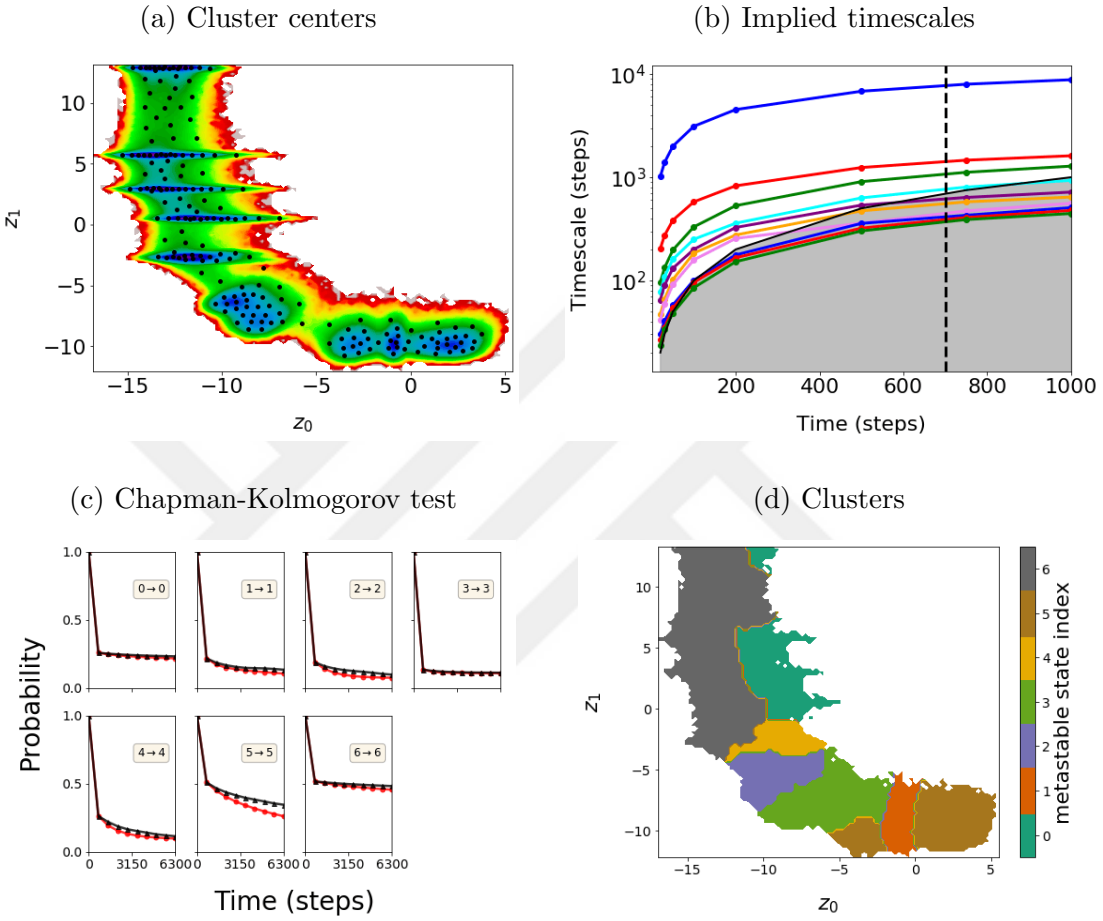


Figure 8.37: Kinetic analysis on the GMVAE landscape for $AAQAA_3$ - I

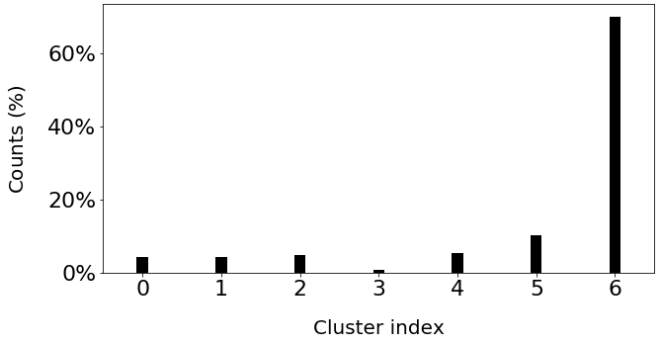


Figure 8.38: Cluster populations for $AAQAA_3$ - I from $PCCA^+$

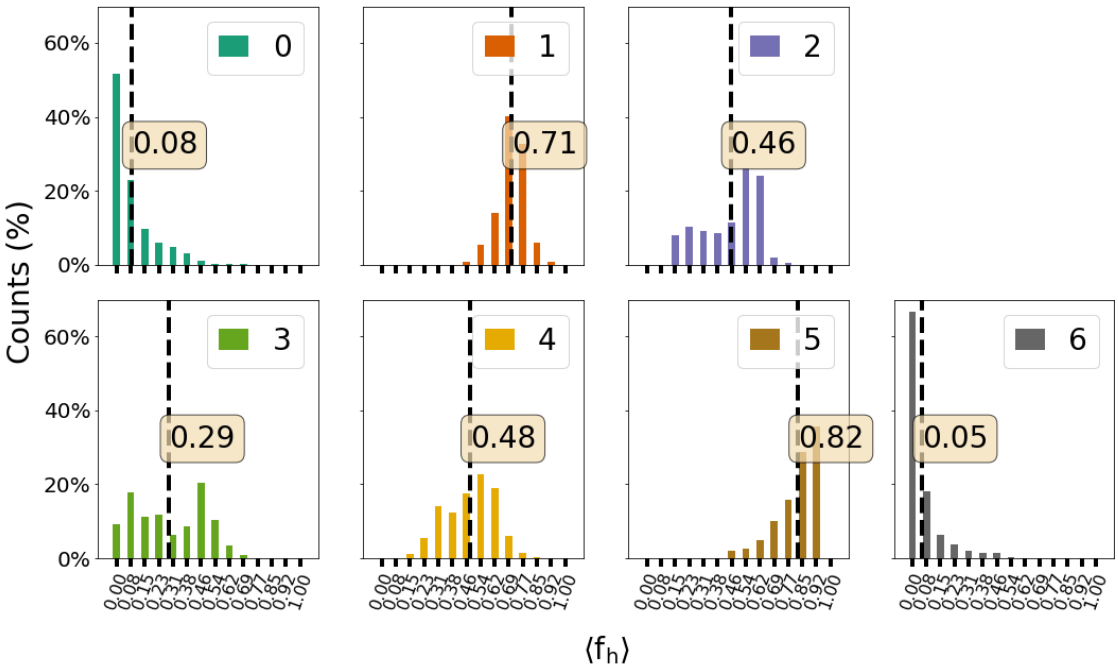


Figure 8.39: Intra-cluster distributions of average helical fraction, $\langle f_h \rangle$, ($AAQAA_3$ - I) for the clusters obtained with PCCA$^+$). The dashed lines indicate the average values, which are also written in the text boxes.

8.3.4.4 TICA results

2D TICA projections are obtained at lag time $\tau = 20$ steps.

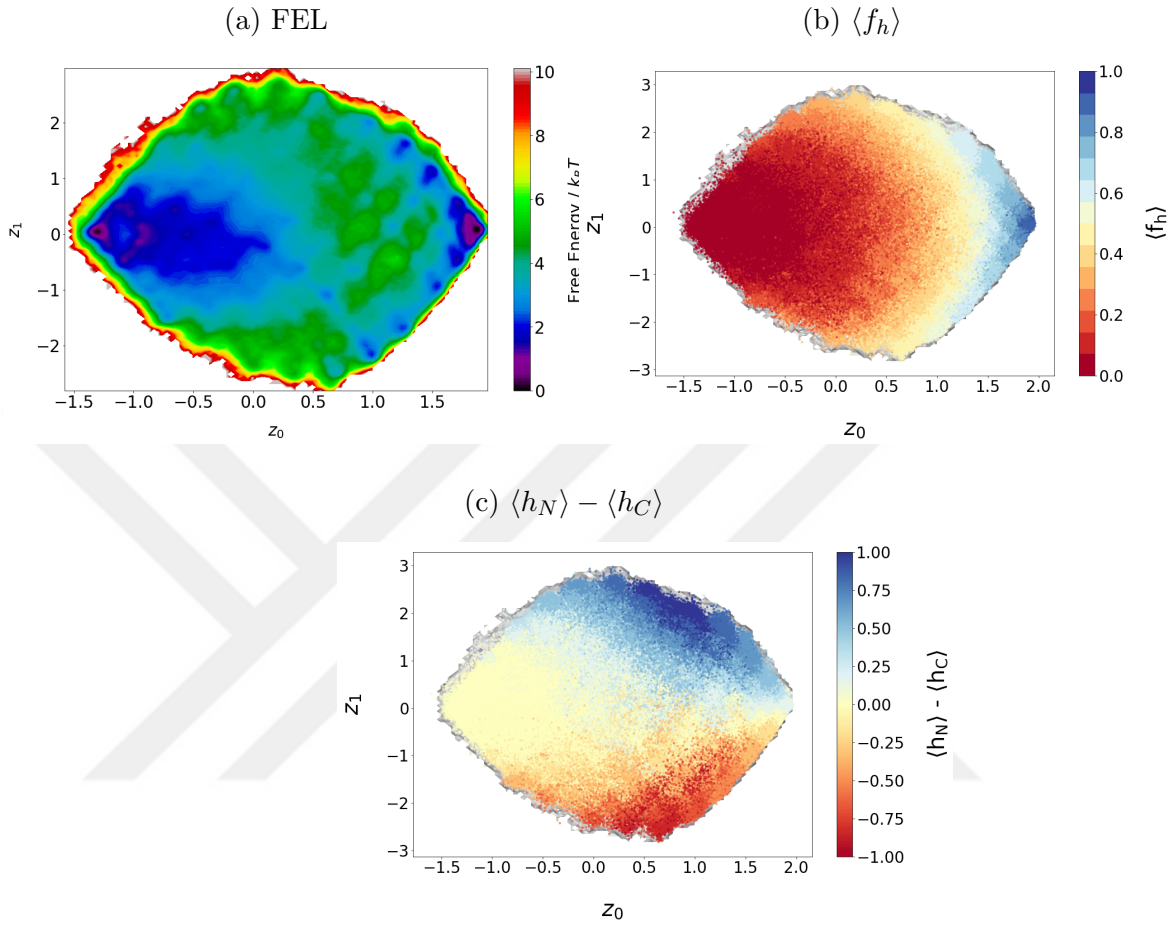


Figure 8.40: TICA results for $AAQAA_3$ - I

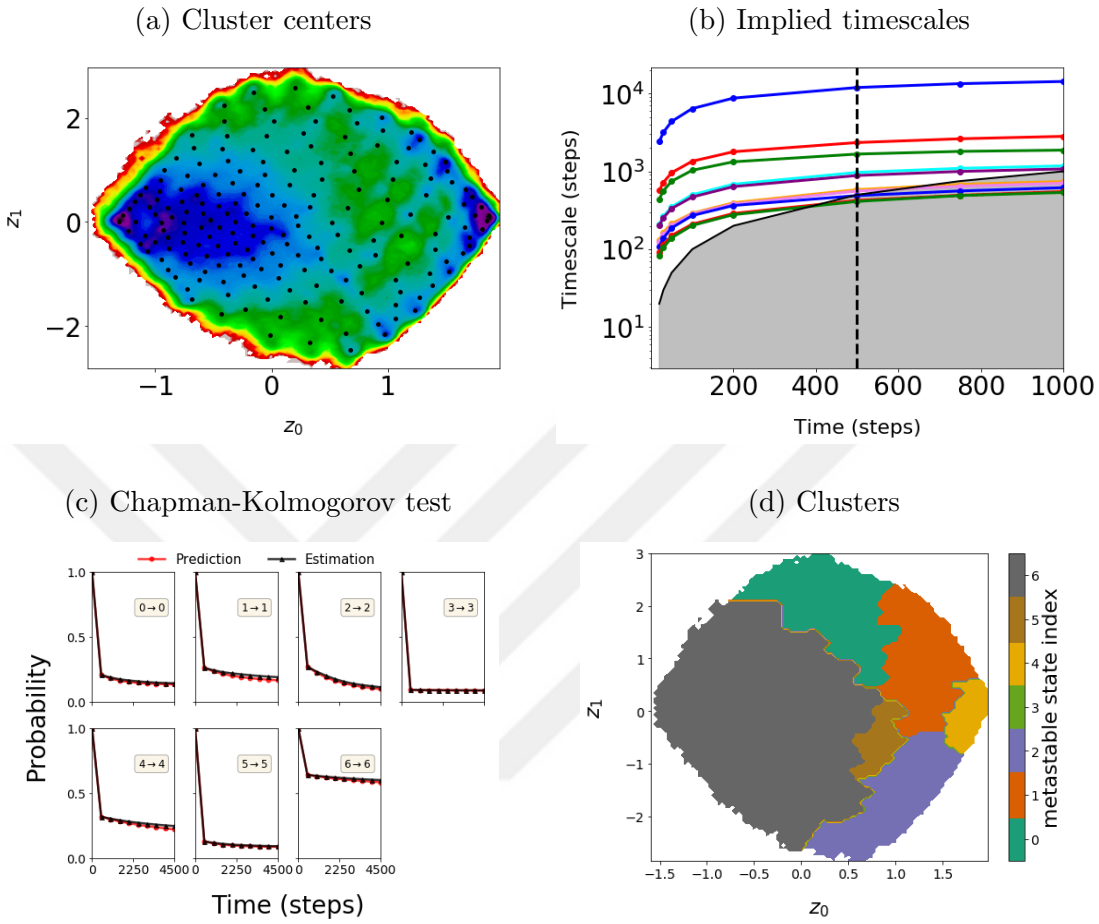


Figure 8.41: Kinetic analysis on TICA landscape for $AAQAA_3$ - I

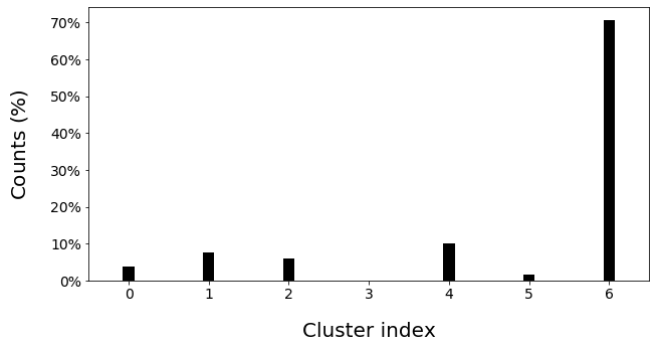


Figure 8.42: Cluster populations for $AAQAA_3$ - I from TICA + PCCA$^+$

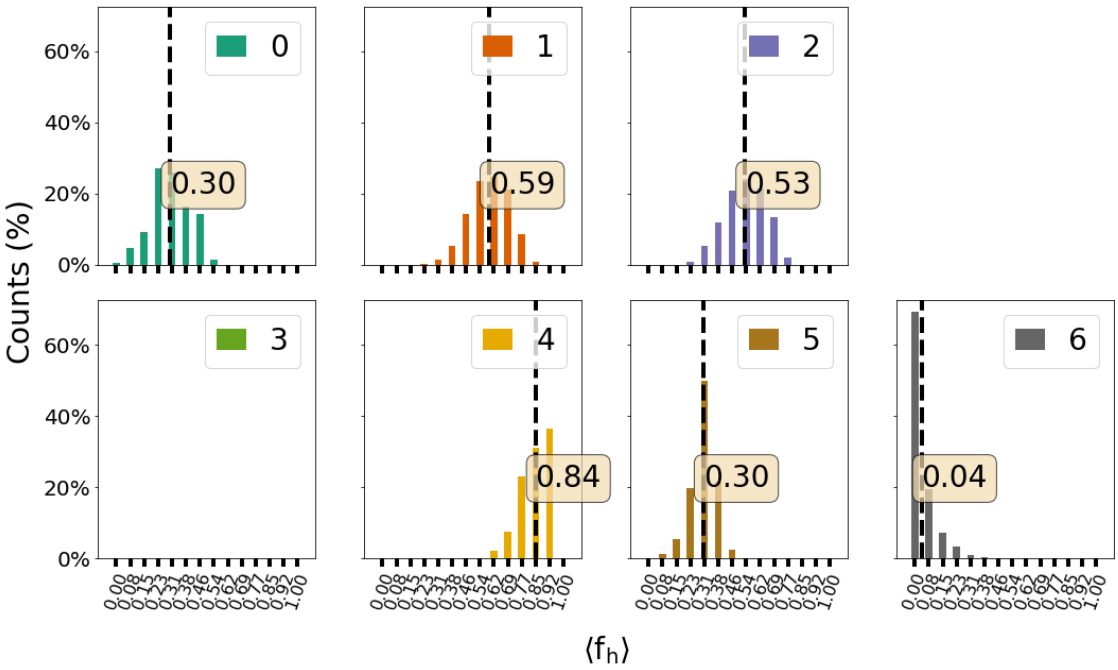


Figure 8.43: Intra-cluster distributions of average helical fraction, $\langle f_h \rangle$, ($AAQAA_3$ - I) for the clusters obtained with PCCA$^+$). The dashed lines indicate the average values, which are also written in the text boxes. Note that cluster 3 is an empty cluster.

8.3.4.5 VAE results

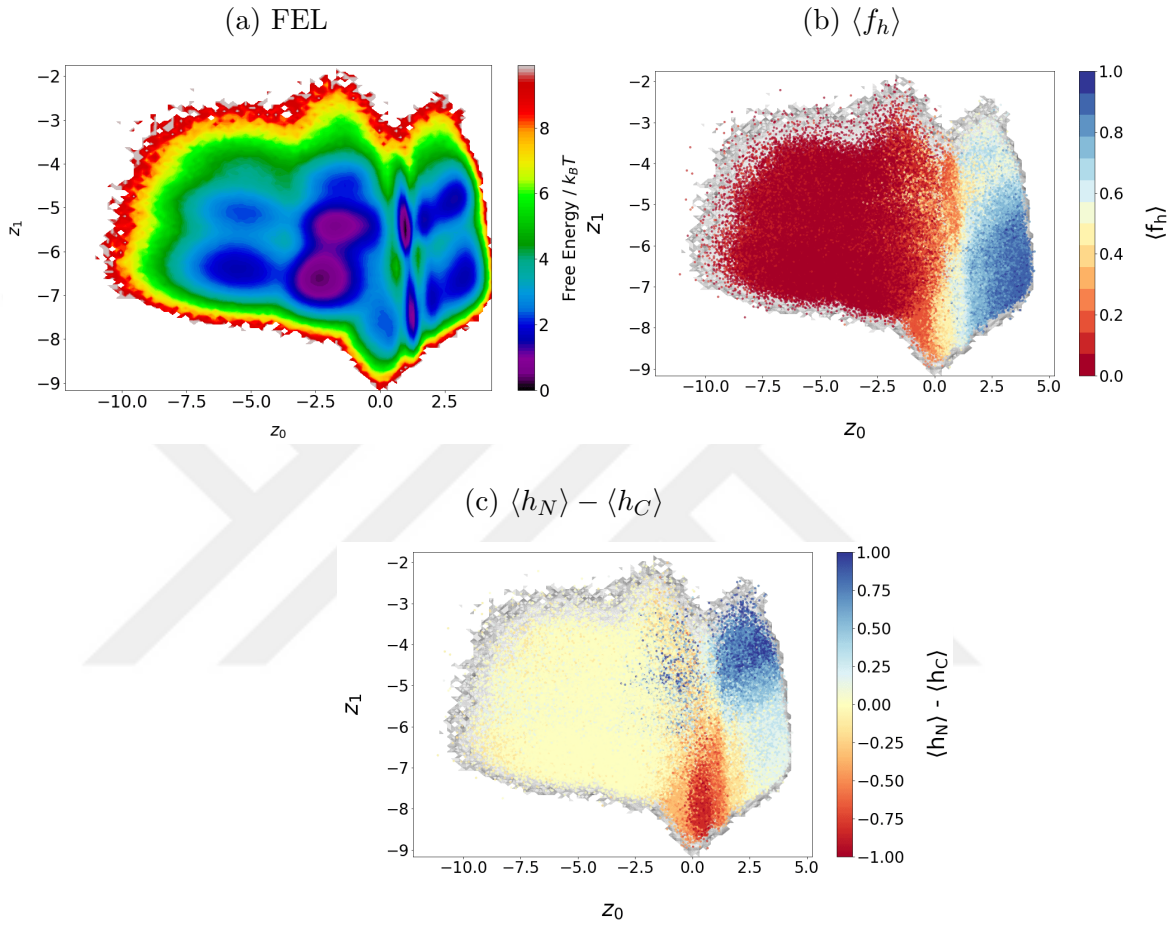


Figure 8.44: VAE results for $AAQAA_3$ - I

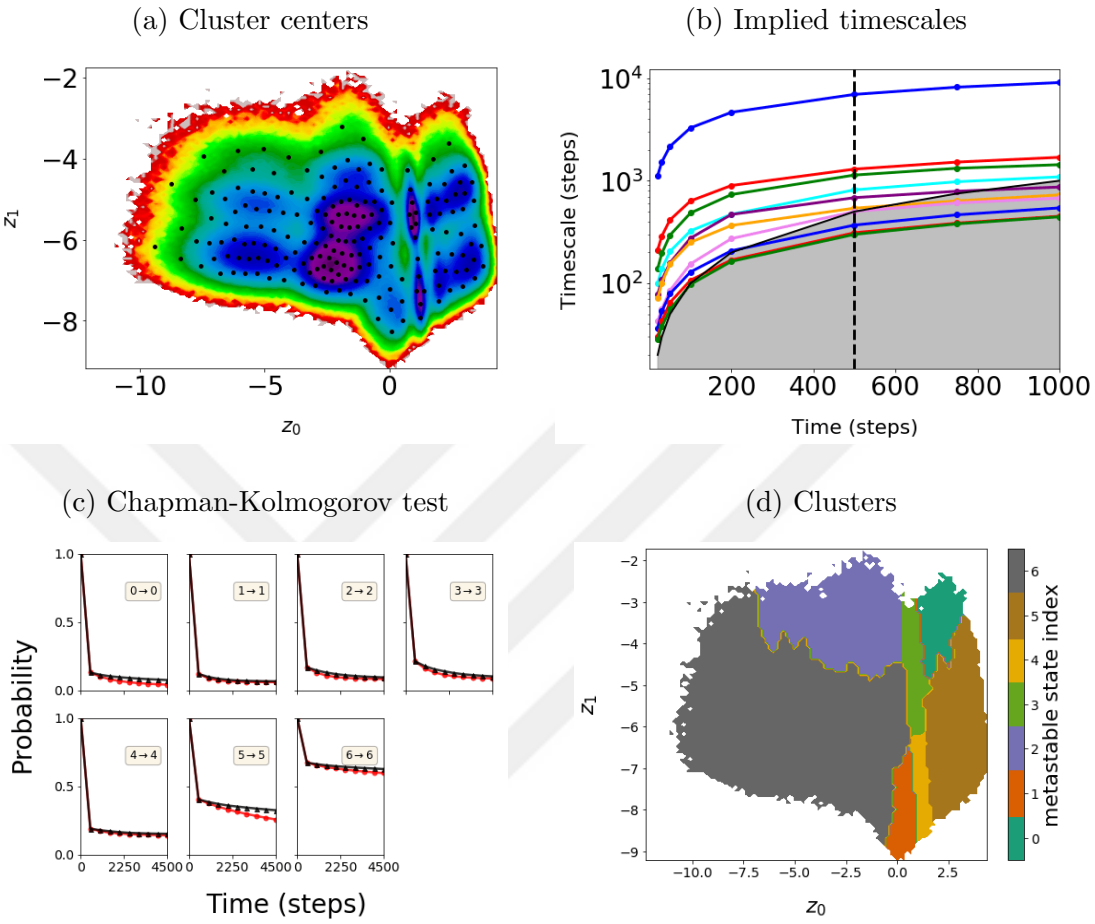


Figure 8.45: Kinetic analysis on the VAE landscape for $AAQAA_3$ - I

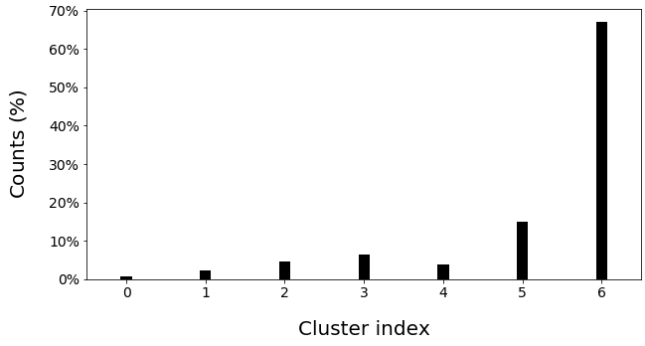


Figure 8.46: Cluster populations for $AAQAA_3$ - I from VAE + PCCA$^+$

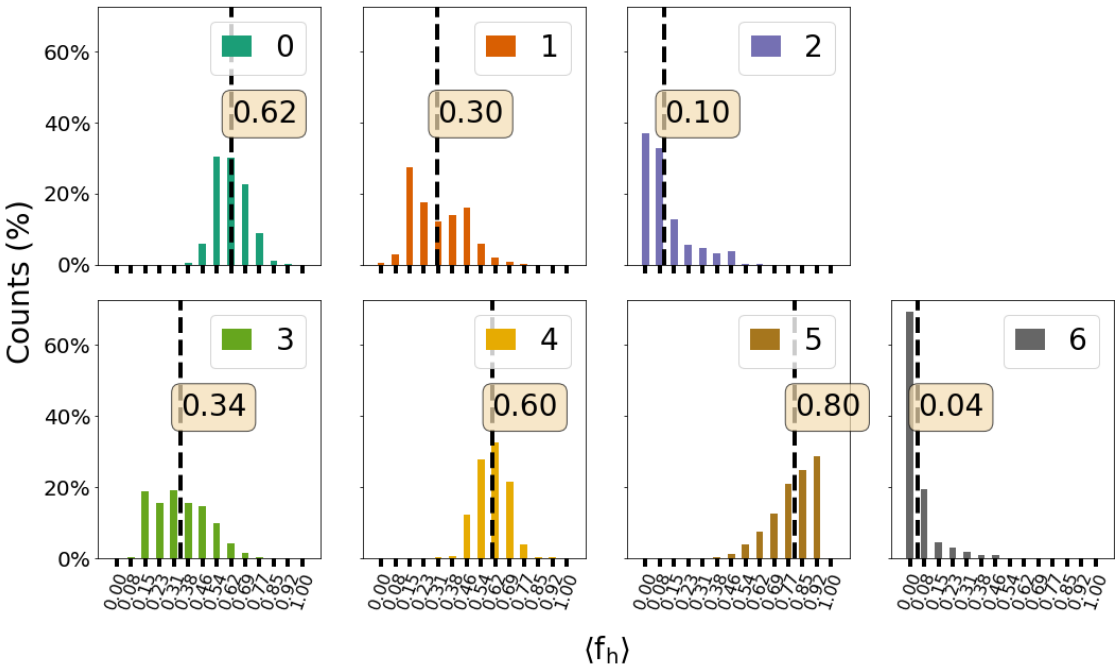


Figure 8.47: Intra-cluster distributions of average helical fraction, $\langle f_h \rangle$, ($AAQAA_3$ - I) for the clusters obtained with PCCA$^+$). The dashed lines indicate the average values, which are also written in the text boxes.

8.3.5 $AAQAA_3$ peptide - II

We also considered an alternative coarse-grained model, with energetic prefactors $\epsilon_{\mathrm{nc}} = 10.92$, $\epsilon_{\mathrm{db}} = 0.2\epsilon_{\mathrm{nc}}$, and $\epsilon_{\mathrm{hp}} = 0.5\epsilon_{\mathrm{nc}}$, while performing simulations at a temperature of 300 K.

8.3.6 GMVAE landscape and the cluster assignments

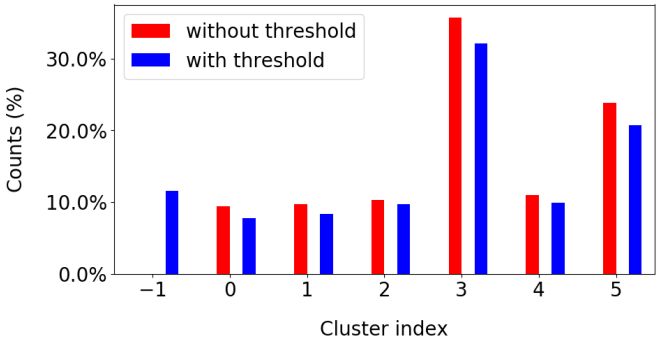


Figure 8.48: Cluster populations from the GMVAE.

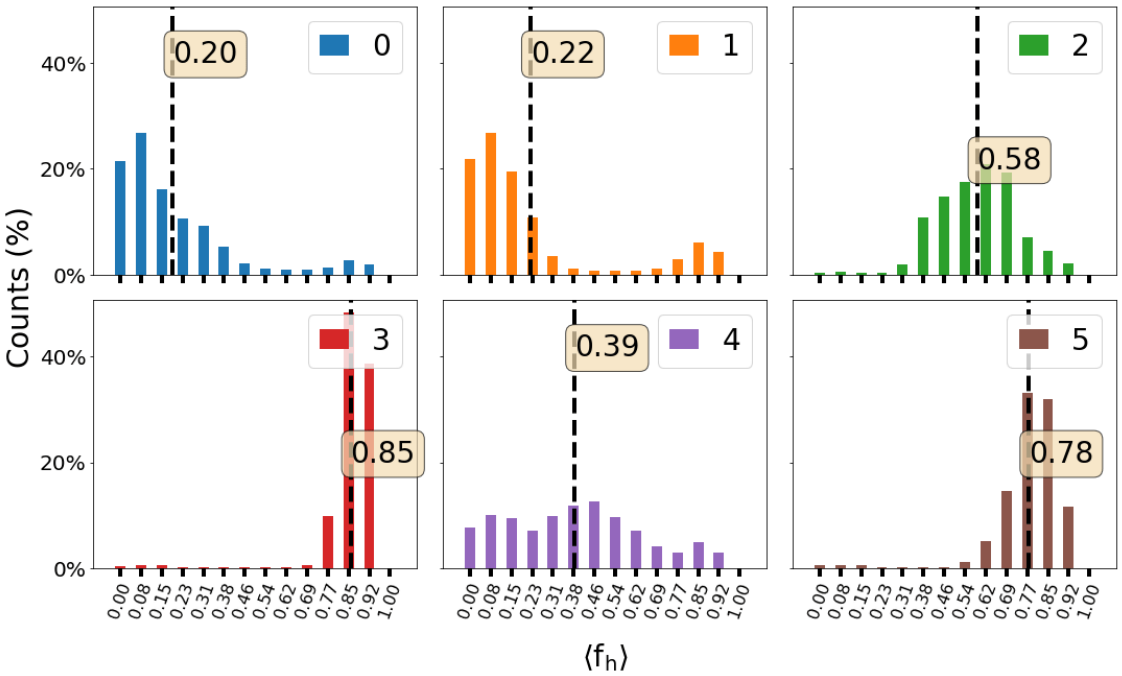


Figure 8.49: Intra-cluster distributions of average helical fraction, $\langle f_h \rangle$, ($AAQAA_3$ - II). The dashed lines indicate the average values, which are also written in the text boxes.

(a) Implied timescales (b) Chapman-Kolmogorov Test

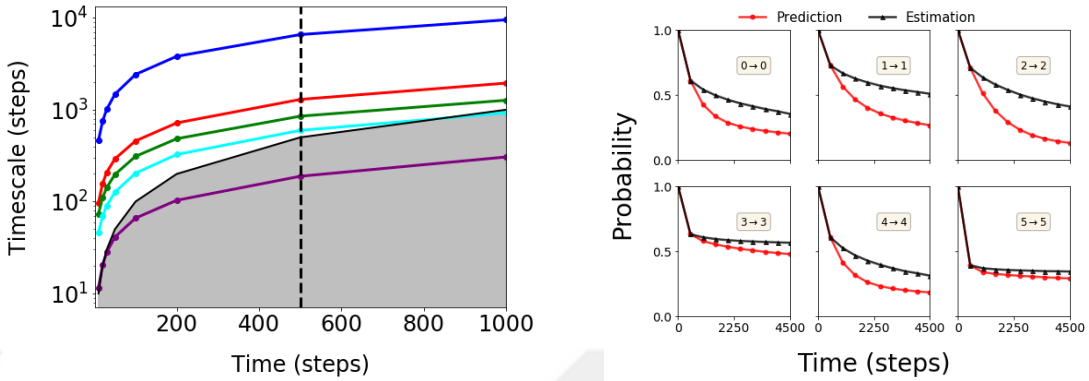


Figure 8.50: Markovianity check of the MSM built for the $AAQAA_3$ - II via using the cluster labels from the GMVAE. (a) Implied timescales. (b) Chapman-Kolmogorov test (at lag=500 steps)

8.3.6.1 GMVAE landscape only (without using the cluster assignments)

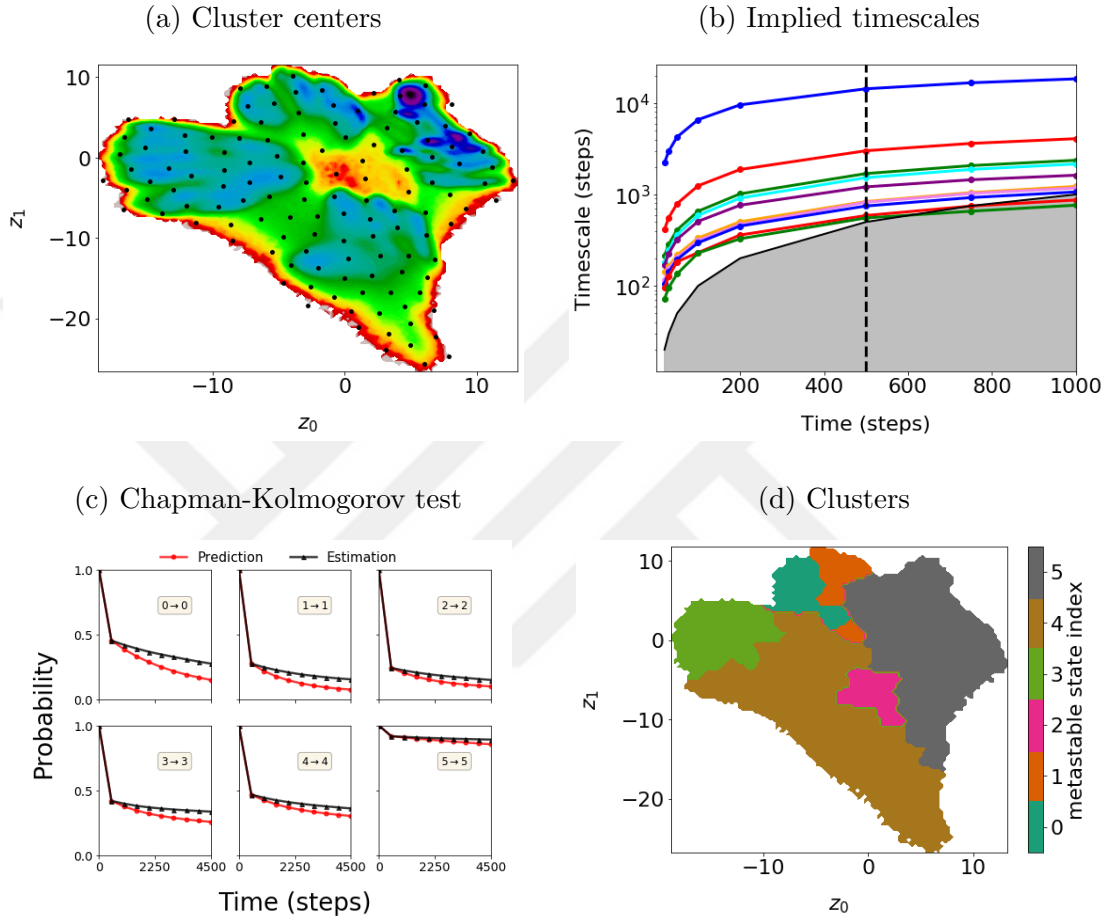


Figure 8.51: Kinetic analysis on the GMVAE landscape for $AAQAA_3$ - II

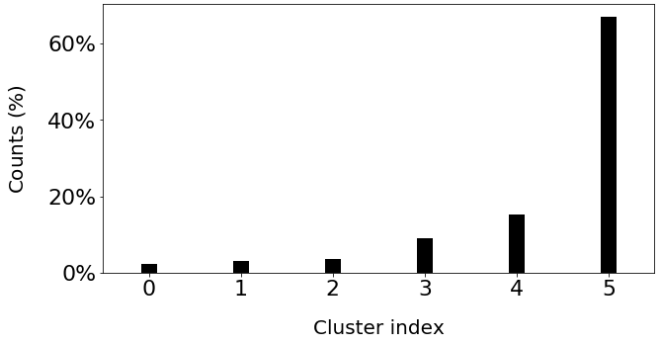


Figure 8.52: Cluster populations for $AAQAA_3$ - II from PCCA$^+$

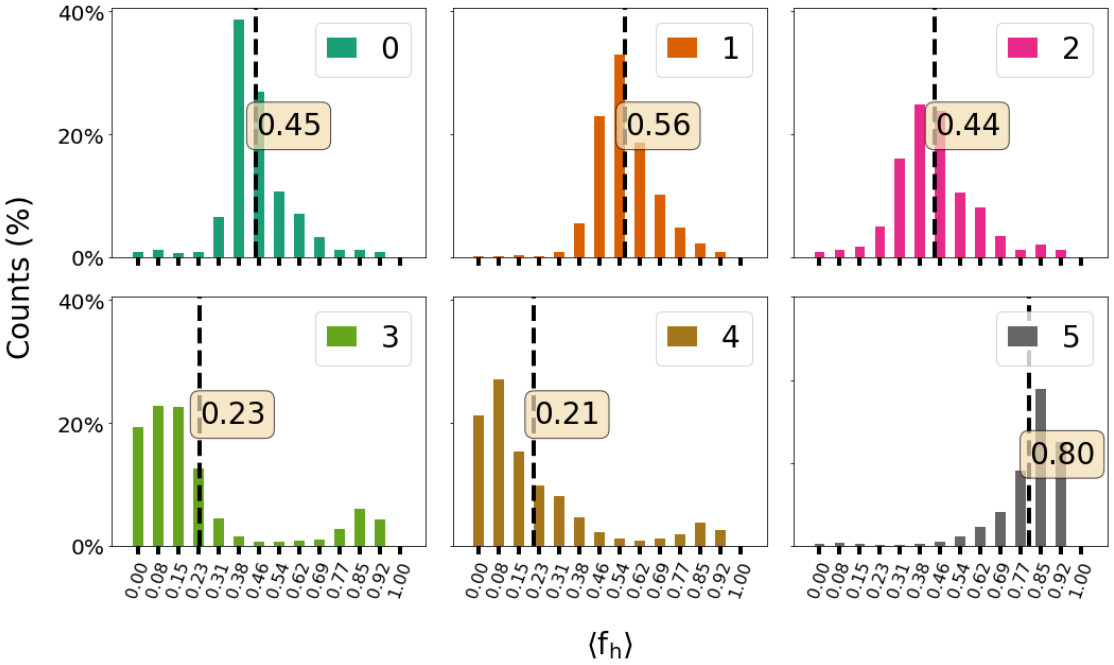


Figure 8.53: Intra-cluster distributions of average helical fraction, $\langle f_h \rangle$, ($AAQAA_3$ - II) for the clusters obtained with $PCCA^+$). The dashed lines indicate the average values, which are also written in the text boxes.

8.3.6.2 TICA results

2D TICA projections are obtained at lag time $\tau = 20$ steps.

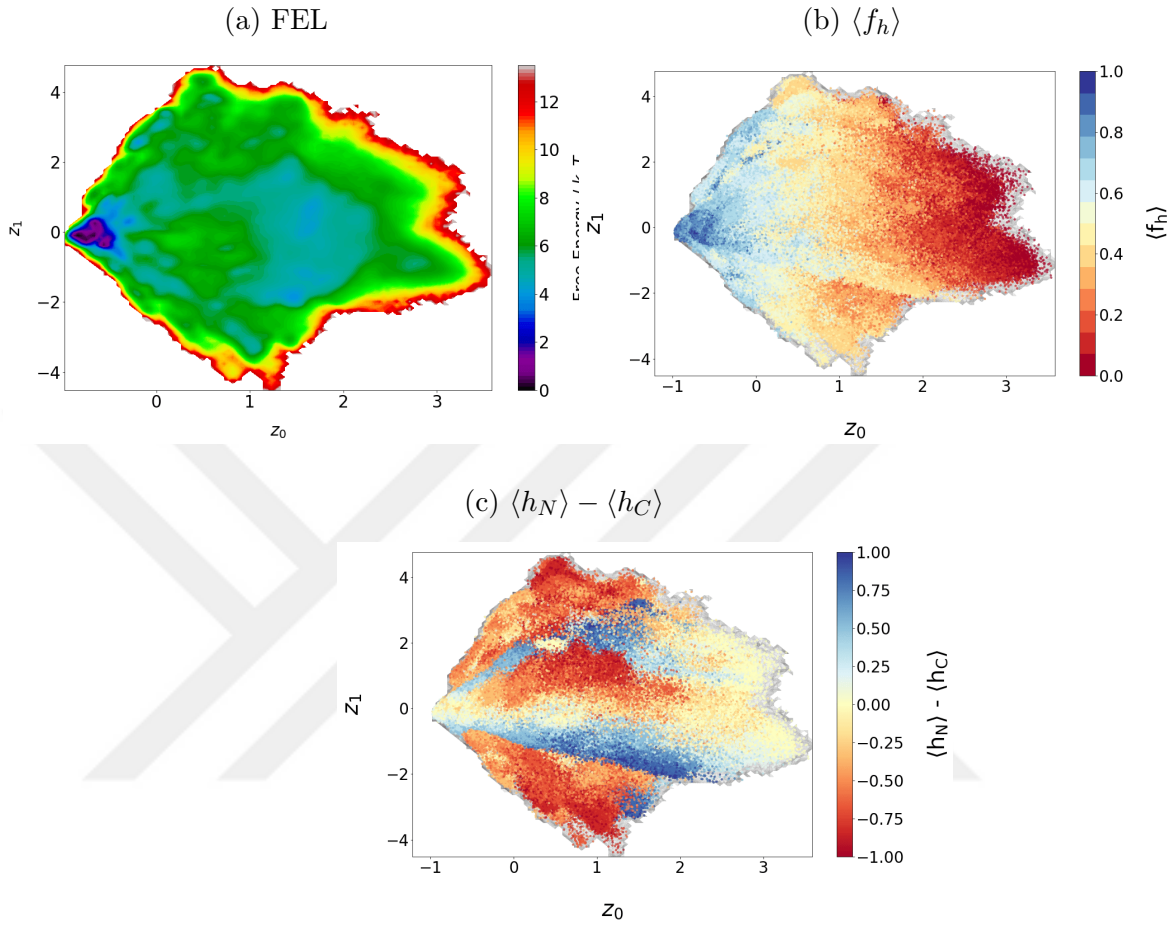


Figure 8.54: TICA results for $AAQAA_3$ - II

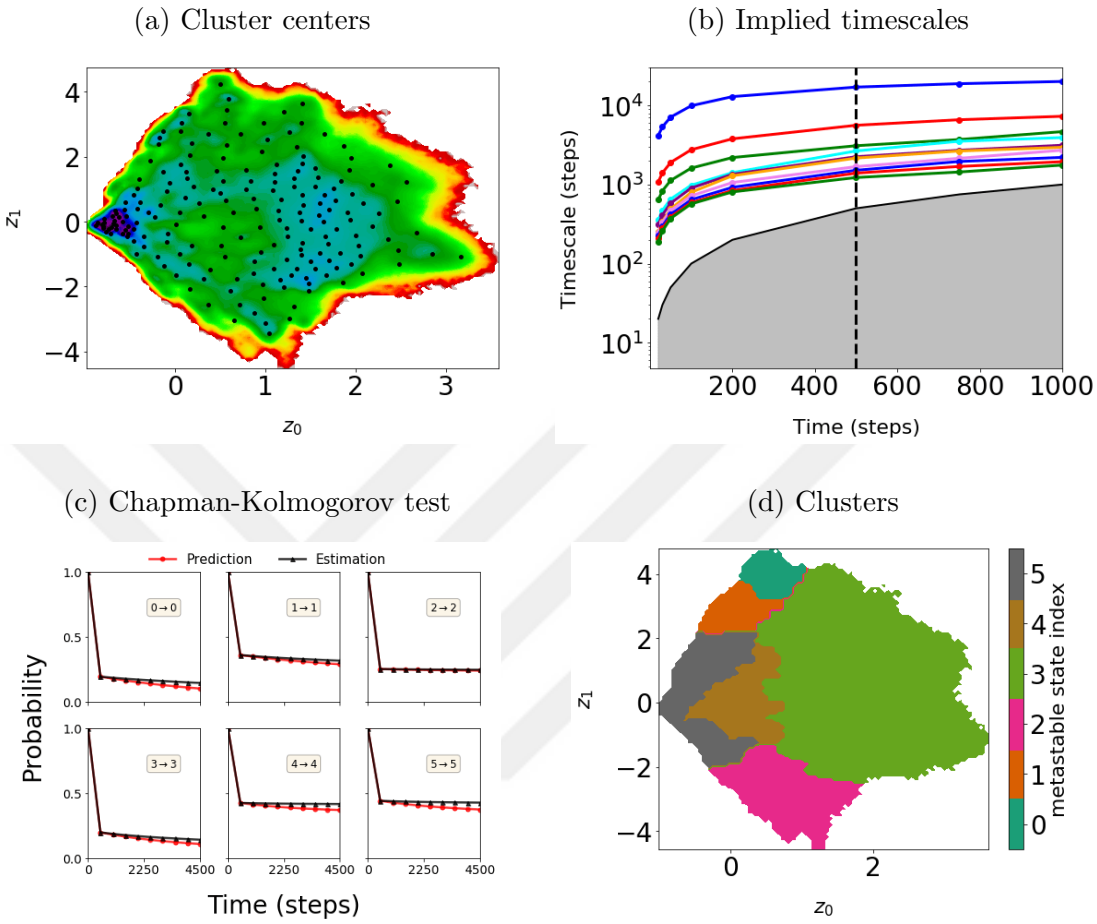


Figure 8.55: Kinetic analysis on TICA landscape for $AAQAA_3$ - II

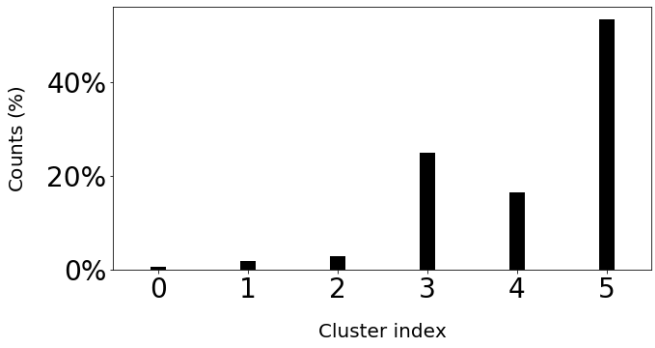


Figure 8.56: Cluster populations for $AAQAA_3$ - II from TICA + PCCA$^+$

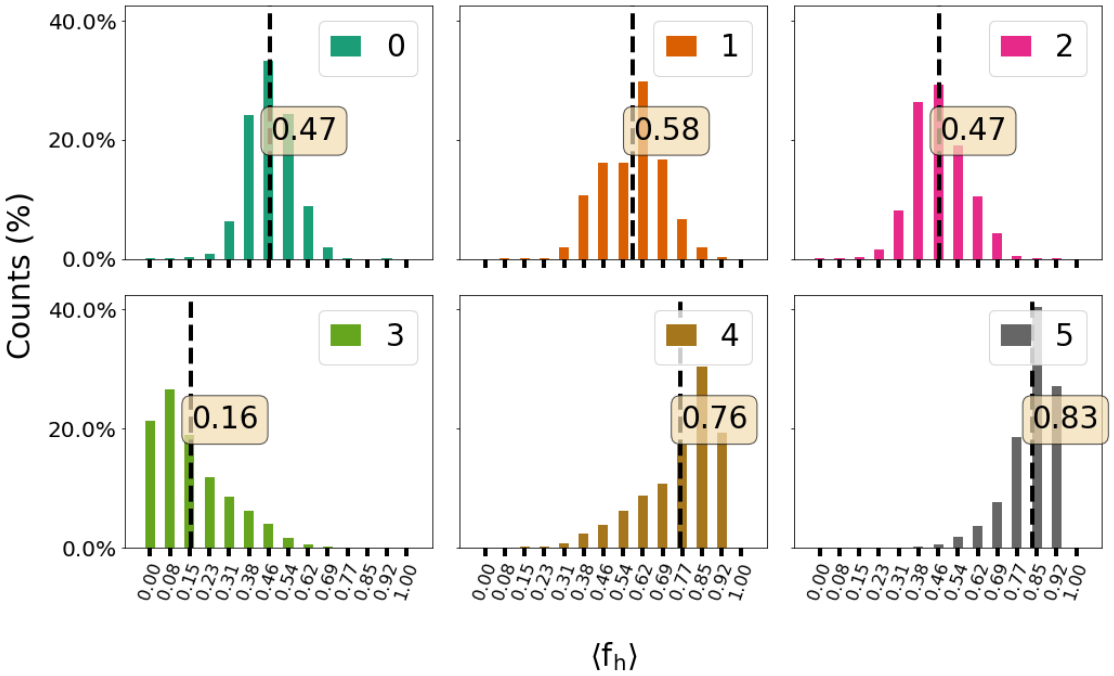


Figure 8.57: Intra-cluster distributions of average helical fraction, $\langle f_h \rangle$, ($AAQAA_3$ - II) for the clusters obtained with PCCA$^+$). The dashed lines indicate the average values, which are also written in the text boxes.

8.3.6.3 VAE results

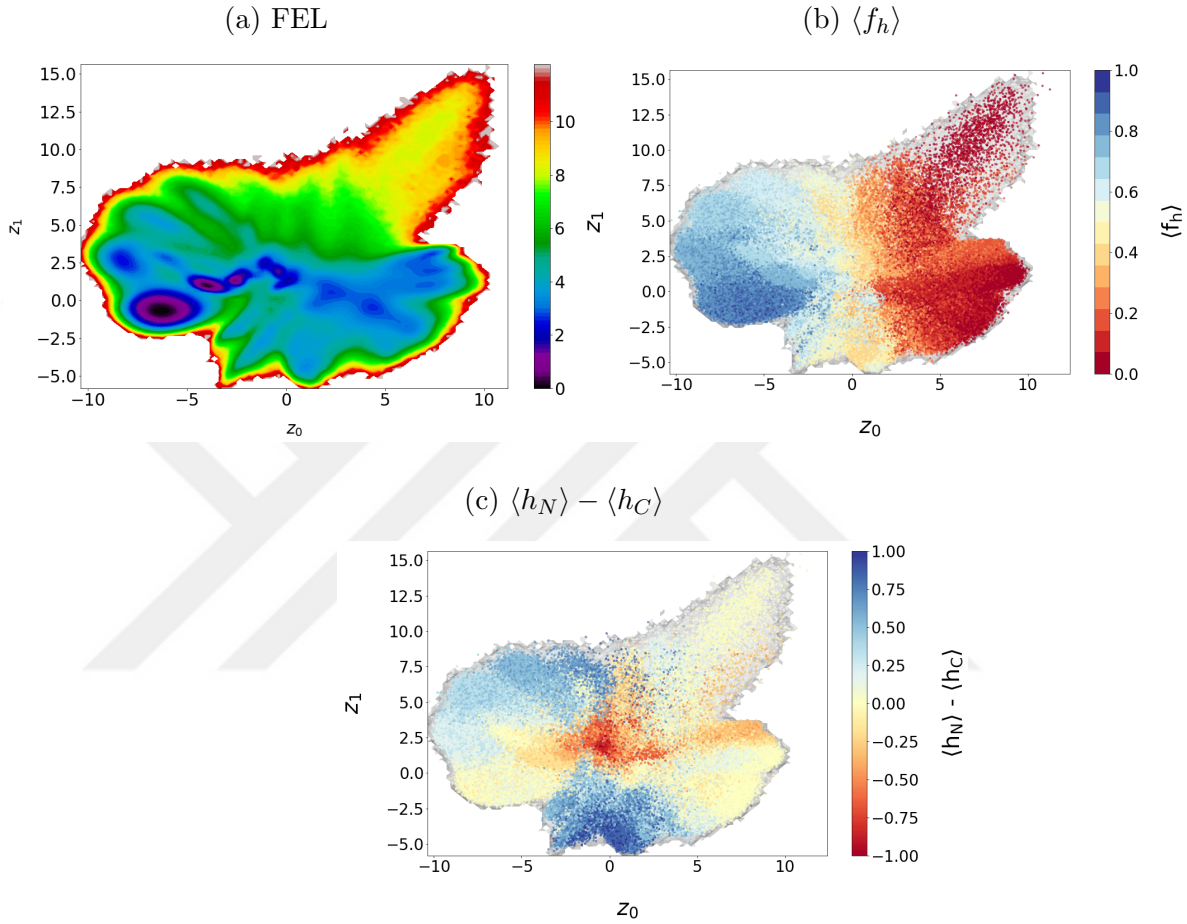


Figure 8.58: VAE results for $AAQAA_3$ - II

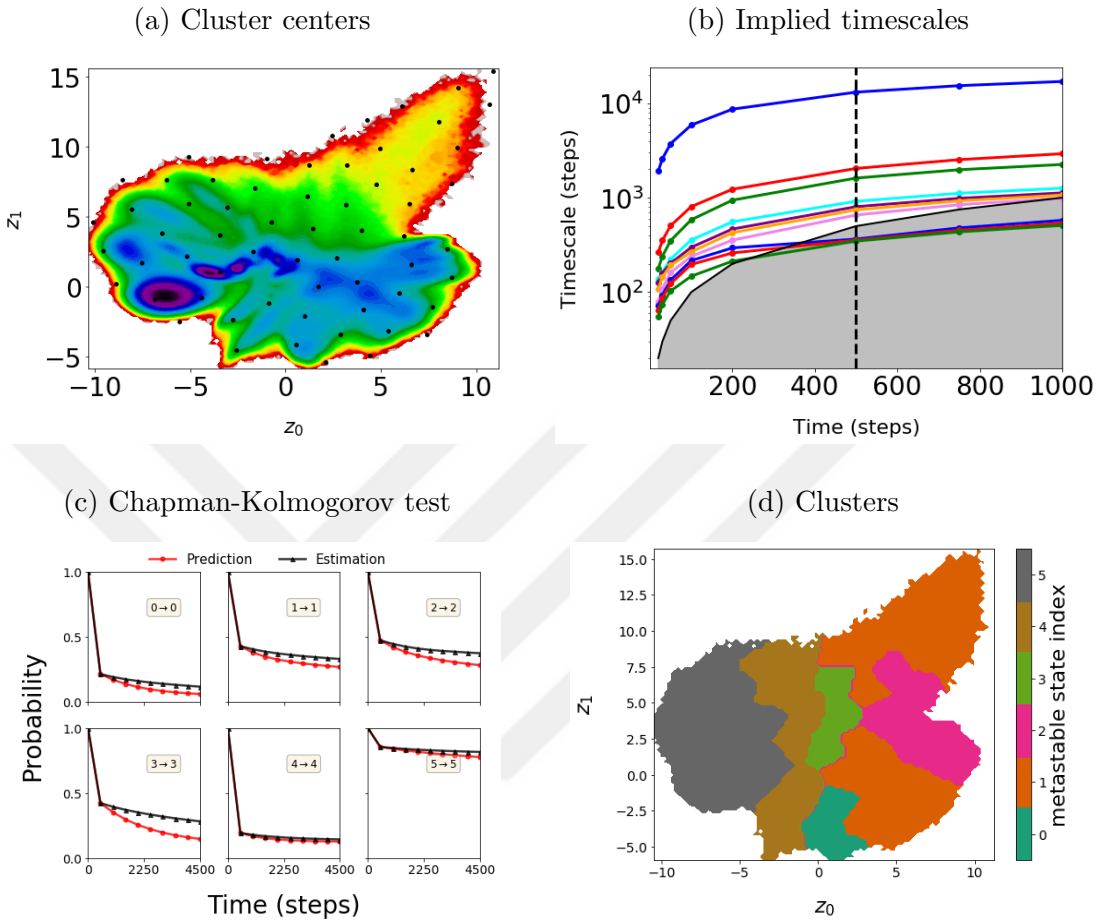


Figure 8.59: Kinetic analysis on the VAE landscape for $AAQAA_3$ - II

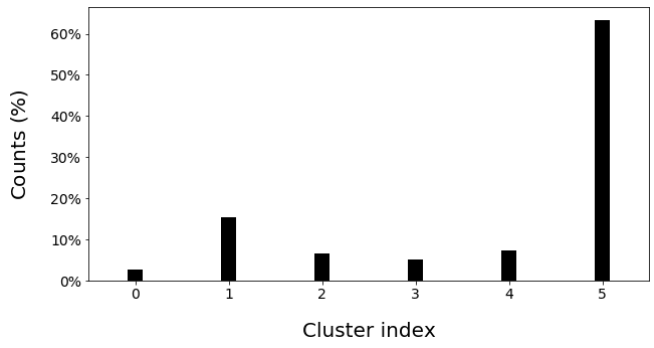


Figure 8.60: Cluster populations for $AAQAA_3$ - II from VAE + PCCA$^+$

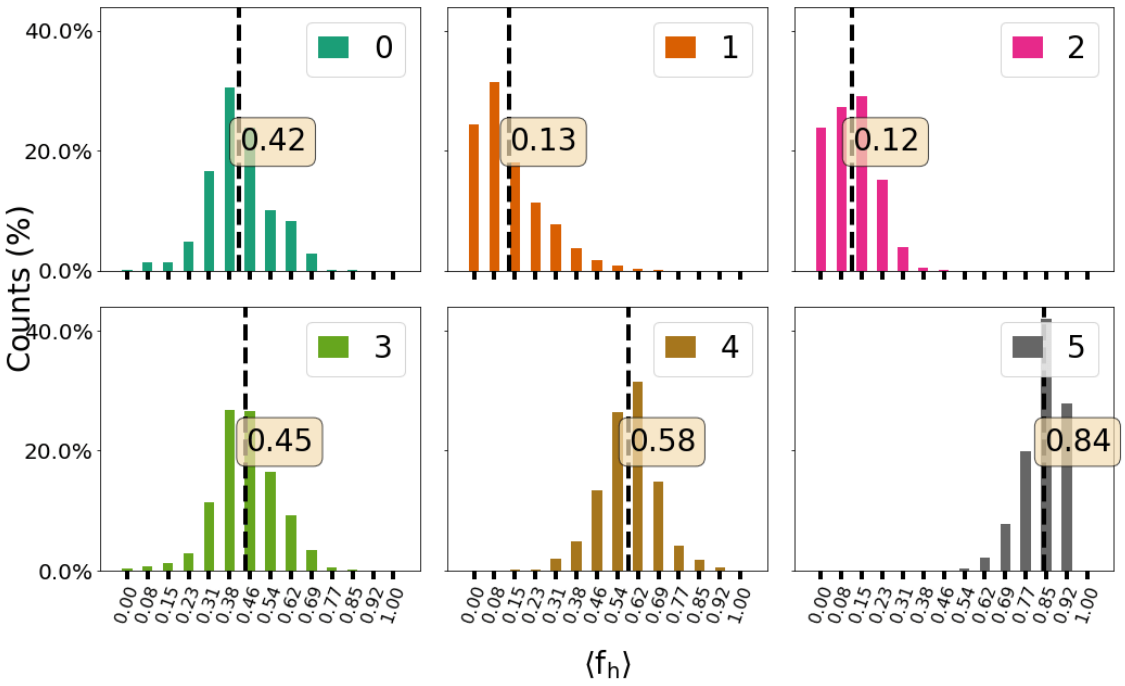


Figure 8.61: Intra-cluster distributions of average helical fraction, $\langle f_h \rangle$, ($AAQAA_3$ - II) for the clusters obtained with PCCA$^+$). The dashed lines indicate the average values, which are also written in the text boxes.

BIBLIOGRAPHY

[1] Broto Chakrabarty and Nita Parekh. Naps: Network analysis of protein structures. *Nucleic acids research*, 44(W1):W375–W382, 2016.

[2] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.

[3] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

[4] Horace R Drew, Richard M Wing, Tsunehiro Takano, Christopher Broka, Shoji Tanaka, Keiichi Itakura, and Richard E Dickerson. Structure of a b-dna dodecamer: conformation and dynamics. *Proceedings of the National Academy of Sciences*, 78(4):2179–2183, 1981.

[5] Z Nevin Gerek and S Banu Ozkan. Change in allosteric network affects binding affinities of pdz domains: analysis through perturbation response scanning. *PLoS Comput Biol*, 7(10):e1002154, 2011.

[6] Celestine N Chi, Lisa Elfström, Yao Shi, Tord Snäll, Åke Engström, and Per Jemth. Reassessing a sparse energetic network within a single protein domain. *Proceedings of the National Academy of Sciences*, 105(12):4679–4684, 2008.

[7] Steve W Lockless and Rama Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, 1999.

[8] Richard N McLaughlin Jr, Frank J Poelwijk, Arjun Raman, Walraj S Gosal, and Rama Ranganathan. The spatial architecture of protein function and adaptation. *Nature*, 491(7422):138, 2012.

[9] Nobuyuki Ota and David A Agard. Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion. *Journal of molecular biology*, 351(2):345–354, 2005.

[10] Wenjun Zheng, Bernard R Brooks, and D Thirumalai. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proceedings of the National Academy of Sciences*, 103(20):7664–7669, 2006.

[11] Bosco K Ho and David A Agard. Conserved tertiary couplings stabilize elements in the pdz fold, leading to characteristic patterns of domain conformational flexibility. *Protein Science*, 19(3):398–411, 2010.

[12] Amit Kumawat and Suman Chakrabarty. Hidden electrostatic basis of dynamic allostery in a pdz domain. *Proceedings of the National Academy of Sciences*, 114(29):E5825–E5834, 2017.

[13] Victor H Salinas and Rama Ranganathan. Coevolution-based inference of amino acid interactions underlying protein function. *Elife*, 7:e34300, 2018.

[14] Stefano Gianni, S Raza Haq, Linda C Montemiglio, Maike C Jürgens, Åke Engström, Celestine N Chi, Maurizio Brunori, and Per Jemth. Sequence-specific long range networks in psd-95/discs large/zo-1 (pdz) domains tune their binding selectivity. *Journal of Biological Chemistry*, 286(31):27167–27175, 2011.

[15] Qi-Shi Du, Cheng-Hua Wang, Si-Ming Liao, and Ri-Bo Huang. Correlation analysis for protein evolutionary family based on amino acid position mutations and application in pdz domain. *PloS one*, 5(10):e13207, 2010.

[16] Robert Kalescky, Hongyu Zhou, Jin Liu, and Peng Tao. Rigid residue scan simulations systematically reveal residue entropic roles in protein allostery. *PLoS computational biology*, 12(4):e1004893, 2016.

[17] Cihan Kaya, Andac Armutlulu, Solen Ekesan, and Turkan Haliloglu. Mcpath: Monte carlo path generation approach to predict likely allosteric pathways and functional residues. *Nucleic acids research*, 41(W1):W249–W255, 2013.

[18] Martin Karplus and J Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural & Molecular Biology*, 9(9):646, 2002.

[19] Peter Deuflhard, Jan Hermans, Benedict Leimkuhler, Alan E Mark, Sebastian Reich, and Robert D Skeel. *Computational Molecular Dynamics: Challenges, Methods, Ideas: Proceeding of the 2nd International Symposium on Algorithms for Macromolecular Modelling, Berlin, May 21–24, 1997*, volume 4. Springer Science & Business Media, 2012.

[20] David E Shaw, Ron O Dror, John K Salmon, JP Grossman, Kenneth M Mackenzie, Joseph A Bank, Cliff Young, Martin M Deneroff, Brannon Batson, Kevin J Bowers, et al. Millisecond-scale molecular dynamics simulations on anton. In *Proceedings of the conference on high performance computing networking, storage and analysis*, pages 1–11, 2009.

[21] John D Chodera and Frank Noé. Markov state models of biomolecular conformational dynamics. *Current opinion in structural biology*, 25:135–144, 2014.

[22] Canan Atilgan and Ali Rana Atilgan. Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. *PLoS Comput Biol*, 5(10):e1000544, 2009.

[23] C Atilgan, ZN Gerek, SB Ozkan, and AR Atilgan. Manipulation of conformational change in proteins by single-residue perturbations. *Biophysical Journal*, 99(3):933–943, 2010.

[24] Kenneth Kundert. *The Designer's Guide to Spice and Spectre®*. Springer Science & Business Media, 2006.

[25] Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*, volume 1. Elsevier, 2001.

[26] Mitsunori Ikeguchi, Jiro Ueno, Miwa Sato, and Akinori Kidera. Protein structural change upon ligand binding: linear response theory. *Physical review letters*, 94(7):078102, 2005.

[27] Alper Demir and Alberto Sangiovanni-Vincentelli. *Analysis and simulation of noise in nonlinear electronic circuits and systems*, volume 425. Springer Science & Business Media, 2012.

[28] Peter Mörters and Yuval Peres. *Brownian motion*, volume 30. Cambridge University Press, 2010.

[29] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[30] Thomas D Schneider. Theory of molecular machines. i. channel capacity of molecular machines. *Journal of theoretical biology*, 148(1):83–123, 1991.

[31] Thomas D Schneider. A brief review of molecular information theory. *Nano communication networks*, 1(3):173–180, 2010.

[32] Ioannis G Kevrekidis, C William Gear, James M Hyman, Panagiotis G Kevrekidid, Olof Runborg, Constantinos Theodoropoulos, et al. Equation-free, coarse-grained multiscale computation: Enabling mocroscopic simulators to perform system-level analysis. *Communications in Mathematical Sciences*, 1(4):715–762, 2003.

[33] Christopher M Dobson. Protein folding and misfolding. *Nature*, 426(6968):884, 2003.

[34] José Nelson Onuchic and Peter G Wolynes. Theory of protein folding. *Current opinion in structural biology*, 14(1):70–75, 2004.

[35] E Weinan, Bjorn Engquist, Xiantao Li, Weiqing Ren, and Eric Vanden-Eijnden. Heterogeneous multiscale methods: a review. *Commun. Comput. Phys*, 2(3):367–450, 2007.

[36] Omar Valsson, Pratyush Tiwary, and Michele Parrinello. Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint. *Annual review of physical chemistry*, 67:159–184, 2016.

[37] Brooke E Husic and Vijay S Pande. Markov state models: From an art to a science. *Journal of the American Chemical Society*, 140(7):2386–2396, 2018.

[38] Wei Chen and Andrew L Ferguson. Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *Journal of computational chemistry*, 39(25):2079–2102, 2018.

[39] João Marcelo Lamim Ribeiro, Pablo Bravo, Yihang Wang, and Pratyush Tiwary. Reweighted autoencoded variational bayes for enhanced sampling (rave). *The Journal of chemical physics*, 149(7):072301, 2018.

[40] Luigi Bonati, Yue-Yu Zhang, and Michele Parrinello. Neural networks based variationally enhanced sampling. *Proceedings of the National Academy of Sciences*, 116(36):17641–17647, 2019.

[41] Ao Ma and Aaron R Dinner. Automatic method for identifying reaction coordinates in complex systems. *The Journal of Physical Chemistry B*, 109(14):6769–6779, 2005.

[42] Philipp Geiger and Christoph Dellago. Neural networks for local structure detection in polymorphic systems. *The Journal of chemical physics*, 139(16):164105, 2013.

[43] Stefan Doerr, Igor Ariz-Extreme, Matthew J Harvey, and Gianni De Fabritiis. Dimensionality reduction methods for molecular simulations. *arXiv preprint arXiv:1710.10629*, 2017.

[44] Tobias Lemke and Christine Peter. Encodermap: Dimensionality reduction and generation of molecule conformations. *Journal of chemical theory and computation*, 15(2):1209–1215, 2019.

[45] Yasemin Bozkurt Varolgunes and Alper Demir. Proteinac: a frequency domain technique for analyzing protein dynamics. *Physical biology*, 15(2):026009, 2018.

[46] Albert Erkip and Burak Erman. Dynamics of large-scale fluctuations in native proteins. analysis based on harmonic inter-residue potentials and random external noise. *Polymer*, 45(2):641–648, 2004.

[47] Abby L Parrill and Kenny B Lipkowitz. Reviews in computational chemistry. volume 28. *Reviews in Computational Chemistry*, 2015.

[48] Ken A Dill, Hue Sun Chan, et al. From levinthal to pathways to funnels. *Nature structural biology*, 4(1):10–19, 1997.

[49] Lars Skjaerven, Siv M Hollup, and Nathalie Reuter. Normal mode analysis for proteins. *Journal of Molecular Structure: THEOCHEM*, 898(1):42–48, 2009.

[50] Osamu Miyashita and Florence Tama. Normal mode analysis techniques in structural biology. *eLS*, 2007.

[51] LT Pillage, RA Rohrer, and C Visweswariah. *Electronic Circuit and System Simulation Methods [Book Reviews]*. McGraw-Hill, Inc., 1995.

[52] Burak Erman. The gaussian network model: precise predictions of residue fluctuations and application to binding problems. *Biophysical journal*, 91(10):3589–3599, 2006.

[53] AR Atilgan, SR Durell, RL Jernigan, MC Demirel, O Keskin, and I Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal*, 80(1):505–515, 2001.

[54] Monique M Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical review letters*, 77(9):1905, 1996.

[55] Eran Eyal and Ivet Bahar. Toward a molecular understanding of the anisotropic response of proteins to external forces: insights from elastic network models. *Biophysical Journal*, 94(9):3424–3435, 2008.

[56] Richard P Feynman, Robert B Leighton, and Matthew Sands. The feynman lectures on physics. *American Journal of Physics*, 33(9):750–752, 1965.

[57] Alper Demir and Burak Erman. Simulation of temporal stochastic phenomena in electronic and biological systems: a comparative review, examples and synergies. In *2013 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 811–818. IEEE, 2013.

[58] Qiang Cui and Ivet Bahar. *Normal mode analysis: theory and applications to biological and chemical systems*. CRC press, 2005.

[59] Martin A Lema and Julian Echave. Assessing local structural perturbations in proteins. *BMC bioinformatics*, 6(1):1, 2005.

[60] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[61] MATLAB. *version 9.4.0 (R2018a)*. The MathWorks Inc., Natick, Massachusetts, 2018.

[62] *The Protein Data Bank*, 2000 (accessed September 4, 2016). `http:www.rcsb.org`.

[63] Yves-Henri Sanejouand. Elastic network models: theoretical and empirical foundations. *Biomolecular Simulations: Methods and Protocols*, pages 601–616, 2013.

[64] Canan Atilgan, Ibrahim Inanc, and Ali Rana Atilgan. Residue network construction and predictions of elastic network models. *arXiv preprint arXiv:0809.3715*, 2008.

[65] TZ Sen and RL Jernigan. Optimizing the parameters of the gaussian network model for atp-binding proteins. *Normal mode analysis: theory and applications to biological and chemical systems*, pages 171–186, 2006.

[66] *Amino Acid Residue Masses*, 2017 (accessed April 6, 2017). `http://www.its.caltech.edu/~ppmal/sample_prep/work3.html`.

[67] Ozge Sensoy, Ali Rana Atilgan, and Canan Atilgan. Fbpa iron storage and release are governed by periplasmic microenvironments. *Physical Chemistry Chemical Physics*, 19(8):6064–6075, 2017.

[68] James E Cassat and Eric P Skaar. Iron in infection and immunity. *Cell host & microbe*, 13(5):509–519, 2013.

[69] Christopher M Bruns, Andrew J Nowalkz, Andrew S Arvaii, Michele A McTigue, Kevin G Vaughanz, Timothy A Mietzner, and Duncan E McRee. evolution within a superfamily. *Nature structural biology*, 4(11):919, 1997.

[70] Karla D Krewulak and Hans J Vogel. Structural biology of bacterial iron uptake. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1778(9):1781–1804, 2008.

[71] Stephen R Shouldice, Douglas R Dougan, Robert J Skene, Leslie W Tari, Duncan E McRee, Rong-hua Yu, and Anthony B Schryvers. High resolution structure of an alternate form of the ferric ion binding protein from haemophilus influenzae. *Journal of Biological Chemistry*, 278(13):11513–11519, 2003.

[72] Simon C Andrews, Andrea K Robinson, and Francisco Rodriguez-Quiñones. Bacterial iron homeostasis. *FEMS microbiology reviews*, 27(2-3):215–237, 2003.

[73] Hakim Boukhalfa, Damon S Anderson, Timothy A Mietzner, and Alvin L Crumbliss. Kinetics and mechanism of iron release from the bacterial ferric binding protein nfbp: exogenous anion influence and comparison with mammalian transferrin. *JBIC Journal of Biological Inorganic Chemistry*, 8(8):881–892, 2003.

[74] Discovery Studio. version 4.1. *Dassault Systemes Biovia Corp. San Diego, CA, USA*, 2014.

[75] Alexander S Rose, Anthony R Bradley, Yana Valasatava, Jose M Duarte, Andreas Prlić, and Peter W Rose. Web-based molecular graphics for large complexes. In *Proceedings of the 21st International Conference on Web3D Technology*, pages 185–186. ACM, 2016.

[76] Alexander S Rose and Peter W Hildebrand. Ngl viewer: a web application for molecular visualization. *Nucleic acids research*, page gkv402, 2015.

[77] Lluiś Bellsolell, Jesús Prieto, Luis Serrano, and Miquel Coll. Magnesium binding to the bacterial chemotaxis protein chey results in large conformational changes involving its functional surface. *Journal of molecular biology*, 238(4):489–495, 1994.

[78] Benjamin Webb and Andrej Sali. Protein structure modeling with modeller. *Protein Structure Prediction*, pages 1–15, 2014.

[79] Mark S Formaneck, Liang Ma, and Qiang Cui. Reconciling the "old" and "new" views of protein allostery: a molecular simulation study of chemotaxis y protein (chey). *PROTEINS: Structure, Function, and Bioinformatics*, 63(4):846–867, 2006.

[80] Ruth Nussinov, Chung-Jung Tsai, and Buyong Ma. The underappreciated role of allostery in the cellular network. *Annual review of biophysics*, 42:169–189, 2013.

[81] Shoshana J Wodak, Emanuele Paci, Nikolay V Dokholyan, Igor N Berezovsky, Amnon Horovitz, Jing Li, Vincent J Hilser, Ivet Bahar, John Karanicolas, Gerhard Stock, et al. Allostery in its many disguises: from theory to applications. *Structure*, 2019.

[82] Jean-Pierre Changeux. The feedback control mechanism of biosynthetic l-threonine deaminase by l-isoleucine. In *Cold Spring Harbor symposia on quantitative biology*, volume 26, pages 313–318. Cold Spring Harbor Laboratory Press, 1961.

[83] Jin Liu and Ruth Nussinov. Allostery: an overview of its history, concepts, methods, and applications. *PLoS computational biology*, 12(6), 2016.

[84] Jacque Monod, Jeffries Wyman, and Jean-Pierre Changeux. On the nature of allosteric transitions: a plausible model. *J Mol Biol*, 12(1):88–118, 1965.

[85] DE Koshland Jr, G Nemethy, and D_ Filmer. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry*, 5(1):365–385, 1966.

[86] Sandeep Kumar, Buyong Ma, Chung-Jung Tsai, Neeti Sinha, and Ruth Nussinov. Folding and binding cascades: dynamic landscapes and population shifts. *Protein science*, 9(1):10–19, 2000.

[87] A Cooper and DTF Dryden. Allostery without conformational change. *European Biophysics Journal*, 11(2):103–109, 1984.

[88] K Gunasekaran, Buyong Ma, and Ruth Nussinov. Is allostery an intrinsic property of all dynamic proteins? *Proteins: Structure, Function, and Bioinformatics*, 57(3):433–443, 2004.

[89] Aysima Hacisuleyman, Batu Erman, Albert Erkip, and Burak Erman. Fluctuation, correlation and perturbation-response behavior of nature-made and artificial nanobodies. *bioRxiv*, 2020.

[90] Claude Elwood Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[91] Carl Eckart. Some studies concerning rotating axes and polyatomic molecules. *Physical Review*, 47(7):552, 1935.

[92] Sotaro Fuchigami, Satoshi Omori, Mitsunori Ikeguchi, and Akinori Kidera. Normal mode analysis of protein dynamics in a non-eckart frame. *The Journal of chemical physics*, 132(10):104109, 2010.

[93] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3):173–181, 1997.

[94] Herbert B Callen and Theodore A Welton. Irreversibility and generalized noise. *Physical Review*, 83(1):34, 1951.

[95] Rep Kubo. The fluctuation-dissipation theorem. *Reports on progress in physics*, 29(1):255, 1966.

[96] Constrained least squares. `http://https://web.stanford.edu/class/ee103/lectures/constrained-least-squares_slides.pdf`. Accessed: 2018-04-11.

[97] Erhan Cinlar. *Introduction to stochastic processes*. Courier Corporation, 2013.

[98] Aleksandr Mikhailovich Lyapunov. The general problem of the stability of motion. *International journal of control*, 55(3):531–534, 1992.

[99] Leon Cohen. Convolution, filtering, linear systems, the wiener-khinchin theorem: generalizations. In *Advanced Signal Processing Algorithms, Architectures, and Implementations III*, volume 1770, pages 378–393. International Society for Optics and Photonics, 1992.

[100] William A Gardner. Introduction to random processes with applications to signals and systems. *New York, MacMillan Co., 1986, 447*, 1986.

[101] Oliver Penrose. *Foundations of statistical mechanics: a deductive treatment*. Courier Corporation, 2005.

[102] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theor. Comp.*, 4(3):435–447, 2008.

[103] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Prot. Struct. Func. Bioinfo.*, 65(3):712–725, NOV 15 2006.

[104] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. Shaw. Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Prot. Struct. Func. Bioinfo.*, 78:1950–1958, 2010.

[105] WL Jorgensen, J Chandrasekhar, JD Madura, RW Impey, and ML Klein. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[106] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraajie. LINCS: a linear constraint solver for molecular simulations. *J. Comp. Chem.*, 18:1463–72, 1997.

[107] RW Hockney, SP Goel, and JW Eastwood. Quiet high-resolution computer models of a plasma. *J. Comp. Phys.*, 14(2):148–158, 1974.

[108] M. P. Allen and D. P. Tildesley. *Computer Simulation of Liquids*. Oxford Press, New York, NY USA, 1987.

[109] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An N log(N) method for ewald sums in large systems. *J. Chem. Phys.*, 99(12):8345–8348, 1993.

[110] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984.

[111] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

[112] Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

[113] Gauss-legendre quadrature. `https://docs.scipy.org/doc/numpy/reference/generated/numpy.polynomial.legendre.leggauss.html`. Accessed: 2020-02-15.

[114] PH D Keith Wilson, Keith Wilson, and John Walker. *Principles and techniques of practical biochemistry*. Cambridge University Press, 2000.

[115] Edwin Joseph Cohn, John Tileston Edsall, et al. Proteins, amino acids and peptides as ions and dipolar ions. 1943.

[116] Candice Gautier, Louise Laursen, Per Jemth, and Stefano Gianni. Seeking allosteric networks in pdz domains. *Protein Engineering, Design and Selection*, 31(10):367–373, 2018.

[117] Chad M Petit, Jun Zhang, Paul J Sapienza, Ernesto J Fuentes, and Andrew L Lee. Hidden dynamic allostery in a pdz domain. *Proceedings of the National Academy of Sciences*, 106(43):18249–18254, 2009.

[118] Yasemin Bozkurt Varolgunes, Tristan Bereau, and Joseph F Rudzinski. Interpretable embeddings from molecular simulations using gaussian mixture variational autoencoders. *Machine Learning: Science and Technology*, 2020.

[119] Kurt Binder. *Monte Carlo and molecular dynamics simulations in polymer science*. Oxford University Press, 1995.

[120] Sandro Bottaro and Kresten Lindorff-Larsen. Biophysical experiments and biomolecular simulations: A perfect match? *Science*, 361(6400):355–360, 2018.

[121] Richard E Bellman. *Adaptive control processes: a guided tour*, volume 2045. Princeton university press, 2015.

[122] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[123] Lutz Molgedey and Heinz Georg Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical review letters*, 72(23):3634, 1994.

[124] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[125] Mary A Rohrdanz, Wenwei Zheng, Mauro Maggioni, and Cecilia Clementi. Determination of reaction coordinates via locally scaled diffusion map. *The Journal of chemical physics*, 134(12):03B624, 2011.

[126] Michele Ceriotti, Gareth A Tribello, and Michele Parrinello. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proceedings of the National Academy of Sciences*, 108(32):13023–13028, 2011.

[127] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[128] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[129] Tameem Adel, Zoubin Ghahramani, and Adrian Weller. Discovering interpretable representations for both deep generative and discriminative models. In *International Conference on Machine Learning*, pages 50–59, 2018.

[130] Christoph Wehmeyer and Frank Noé. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *The Journal of chemical physics*, 148(24):241703, 2018.

[131] Carlos X Hernández, Hannah K Wayment-Steele, Mohammad M Sultan, Brooke E Husic, and Vijay S Pande. Variational encoding of complex dynamics. *Physical Review E*, 97(6):062412, 2018.

[132] Wei Chen, Hythem Sidky, and Andrew L Ferguson. Capabilities and limitations of time-lagged autoencoders for slow mode discovery in dynamical systems. *The Journal of Chemical Physics*, 151(6):064123, 2019.

[133] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

[134] Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. Vampnets for deep learning of molecular kinetics. *Nature communications*, 9(1):1–11, 2018.

[135] Bethany Lusch, J Nathan Kutz, and Steven L Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature communications*, 9(1):4950, 2018.

[136] Variational autoencoder: Intuition and implementation. `https://wiseodd.github.io/techblog/2016/12/10/variational-autoencoder/`. Accessed: 2019-04-10.

[137] Gaussian mixture vae: Lessons in variational inference, generative models, and deep nets. `http://ruishu.io/2016/12/25/gmvae/`. Accessed: 2019-05-14.

[138] Qingyu Zhao, Nicolas Honnorat, Ehsan Adeli, Adolf Pfefferbaum, Edith V Sullivan, and Kilian M Pohl. Variational autoencoder with truncated mixture of gaussians for functional connectivity analysis. In *International Conference on Information Processing in Medical Imaging*, pages 867–879. Springer, 2019.

[139] Eric Nalisnick, Lars Hertel, and Padhraic Smyth. Approximate inference for deep latent gaussian mixtures. In *NIPS Workshop on Bayesian Deep Learning*, volume 2, 2016.

[140] Wenxian Shi, Hao Zhou, Ning Miao, Shenjian Zhao, and Lei Li. Fixing gaussian mixture vaes for interpretable text generation. *arXiv preprint arXiv:1906.06719*, 2019.

[141] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and

Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[142] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[143] Bowman, Gregory R. and Pande, Vijay S. and Noé, Frank. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*. Springer Science and Business Media, Dordrecht, Netherlands, 2014.

[144] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D Chodera, Christof Schütte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *The Journal of chemical physics*, 134(17):174105, 2011.

[145] Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for markov model construction. *The Journal of chemical physics*, 139(1):07B604_1, 2013.

[146] Susanna Röblitz and Marcus Weber. Fuzzy spectral clustering by pcca+: Application to markov state models and data classification. *Advances in Data Analysis and Classification*, 7(2):147–179, 2013.

[147] Martin K Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Perez-Hernandez, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noe. Pyemma 2: A software package for estimation, validation, and analysis of markov models. *Journal of chemical theory and computation*, 11(11):5525–5542, 2015.

[148] Shneior Lifson and A Roig. On the theory of helix coil transition in polypeptides. *The Journal of Chemical Physics*, 34(6):1963–1974, 1961.

[149] Andrew James Doig. The a-helix as the simplest protein model: Helix-coil theory, stability, and design. *Protein Folding, Misfolding and Aggregation*, page 1, 2008.

[150] Debsindhu Bhowmik, Shang Gao, Michael T Young, and Arvind Ramanathan. Deep clustering of protein folding simulations. *BMC bioinformatics*, 19(18):484, 2018.

[151] Christian R Schwantes and Vijay S Pande. Modeling molecular kinetics with tica and the kernel trick. *Journal of chemical theory and computation*, 11(2):600–608, 2015.

[152] Klaus Müller and Leo D Brown. Location of saddle points and minimum energy paths by a constrained simplex optimization procedure. *Theoretica chimica acta*, 53(1):75–93, 1979.

[153] Feliks Nüske, Hao Wu, Jan-Hendrik Prinz, Christoph Wehmeyer, Cecilia Clementi, and Frank Noé. Markov state models from short non-equilibrium simulations: Analysis and correction of estimation bias. *The Journal of Chemical Physics*, 146(9):094104, 2017.

[154] Jun Zhang, Yao-Kun Lei, Xing Che, Zhen Zhang, Yi Isaac Yang, and Yi Qin Gao. Deep representation learning for complex free-energy landscapes. *The journal of physical chemistry letters*, 10(18):5571–5576, 2019.

[155] Wei Chen, Hythem Sidky, and Andrew L Ferguson. Nonlinear discovery of slow molecular modes using state-free reversible vampnets. *The Journal of chemical physics*, 150(21):214114, 2019.

[156] `https://github.com/markovmodel/mdshare`.

[157] Alexandros Altis, Moritz Otten, Phuong H Nguyen, Rainer Hegger, and Gerhard Stock. Construction of the free energy landscape of biomolecules via

dihedral angle principal component analysis. *The Journal of chemical physics*, 128(24):06B620, 2008.

[158] John D Chodera, Nina Singhal, Vijay S Pande, Ken A Dill, and William C Swope. Automatic discovery of metastable states for the construction of markov models of macromolecular conformational dynamics. *The Journal of chemical physics*, 126(15):04B616, 2007.

[159] Joseph F Rudzinski and Tristan Bereau. Structural-kinetic-thermodynamic relationships identified from physics-based molecular simulation models. *The Journal of chemical physics*, 148(20):204111, 2018.

[160] Joseph Rudzinski and Tristan Bereau. The role of conformational entropy in the determination of structural-kinetic relationships for helix-coil transitions. *Computation*, 6(1):21, 2018.

[161] Oleksandra Kukharenko, Kevin Sawade, Jakob Steuer, and Christine Peter. Using dimensionality reduction to systematically expand conformational sampling of intrinsically disordered peptides. *Journal of chemical theory and computation*, 12(10):4726–4734, 2016.

[162] Abhinav Jain and Gerhard Stock. Identifying metastable states of folding proteins. *Journal of chemical theory and computation*, 8(10):3810–3819, 2012.

[163] `https:https://github.com/hsidky/srv`.

[164] `https://github.com/msmbuilder/vde`.

[165] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W.

Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. *arXiv e-prints*, page arXiv:1907.10121, Jul 2019.

[166] Joseph F Rudzinski and Tristan Bereau. The role of conformational entropy in the determination of structural-kinetic relationships for helix-coil transitions. *Computation*, 6(1):21, 2018.

[167] H Taketomi, Y Ueda, and N Gō. Studies on protein folding, unfolding and fluctuations by computer-simulation. 1. effect of specific amino-acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.*, 7(6):445–459, 1975.

[168] MS Cheung, AE Garcia, and JN Onuchic. Protein folding mediated by solvation: Water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc. Natl. Acad. Sci. USA*, 99(2):685–690, 2002.

[169] Cecilia Clementi and Steven S. Plotkin. The effects of nonnative interactions on protein folding rates: Theory and simulation. *Protein Sci.*, 13:1750–1766, 2004.

[170] Hue Sun Chan, Zhuqing Zhang, Stefan Wallin, and Zhirong Liu. Cooperativity, Local-Nonlocal Coupling, and Nonnative Interactions: Principles of Protein Folding from Coarse-Grained Models. *Ann. Rev. Phys. Chem.*, 62:301–326, 2011.