Modelling Allostery using a Computational Analysis of Side-Chain Interactions

by

Leonard Zhao

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2022

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Allostery refers to the regulation of protein activity arising from an effector molecule, such as a ligand, binding to a protein. The exact mechanisms that take place in allosteric regulation are still a source of debate within the protein research community. To gain a better understanding of allosteric mechanisms, we devised a computational model of allostery that focused on simple protein structures that have a fixed backbone but dynamic side-chains.

Our model relied on a statistical analysis of side-chain couplings to determine the effect of side-chain fluctuations. To obtain the side-chain dataset required for the statistical analysis, we used an energy minimization procedure contained within the UCSF Chimera molecular modelling software to evaluate concerted side-chain movements. We also derived residue networks and designed graph algorithms that mimicked allosteric signal propagations. These techniques enabled us to identify highly fluctuating sites within a protein structure and to uncover potential functionally important residues.

We evaluated our methods by applying them to the PDZ3 domain of the PSD-95 protein. This protein structure was chosen due to its relatively small size and rigid backbone. We identified residues that experienced high levels of side-chain fluctuations, and our results agreed with experimentally determined functionally important residues. Comparing the results for the apo and holo forms of the protein also revealed structural elements, such as side-chain fluctuations within alpha helices, that are important for allosteric signal transmissions.

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Forbes J. Burkowski, whose support and guidance made this thesis possible. Doing research in such a broad and interdisciplinary field is never an easy task, but Forbes was always able to steer me in the right direction.

I would also like to thank Professors Bin Ma and Brendan McConkey for agreeing to read this thesis.

Finally, I would like to thank my family and friends for helping me make it through these past two years.

Table of Contents

List of Figures

List of Tables

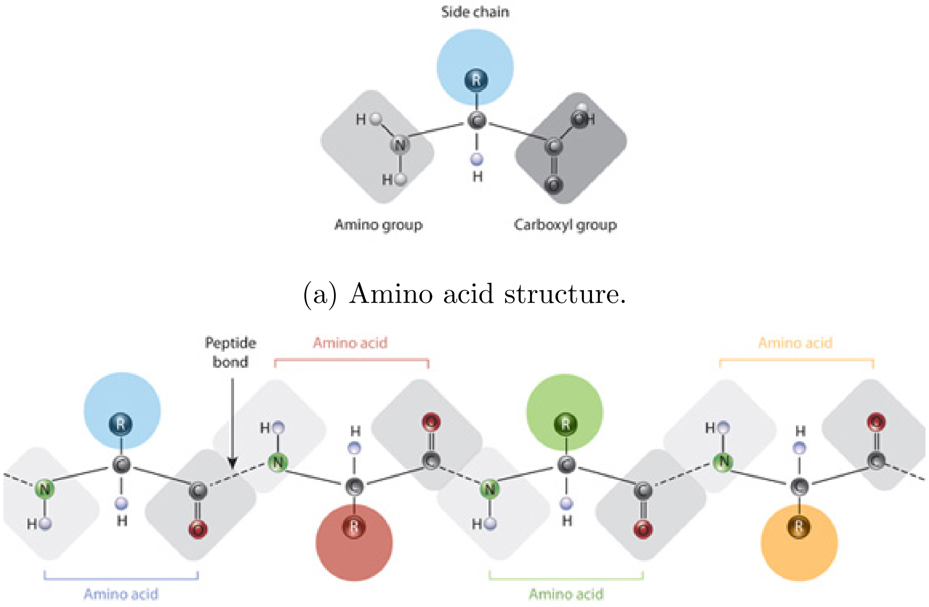
Chapter 1

Introduction

Proteins are a fundamental component of living organisms and are responsible for a vast array of functions. Each protein is composed of a connected sequence of amino acids, which are also called residues. Each amino acid consists of a side-chain group, a carboxyl group, and an amino group. The three-dimensional structure of proteins is composed of a fairly rigid backbone along with more flexible side-chains. The backbone, which includes the sequence of carboxyl and amino groups of residues connected with peptide bonds, holds the protein together and determines the overall three-dimensional shape. Side-chains branch off from the backbone and define each residue type. The composition of a residue's side-chain provides that residue with distinctive molecular properties, such as charge, polarity, and hydrophobicity. Furthermore, the flexibility of side-chains allows for interactions between residues that are physically close within a protein.

Protein structure is typically described in four different levels. The primary structure describes the linear sequence of amino acids that are connected. The secondary structure describes the local three-dimensional configurations within a protein, and includes alpha helices, beta sheets, and loops. Tertiary structure describes the overall three-dimensional shape of the protein. Each atom within the protein can be given a coordinate in $\mathbb{R}^3$, and the arrangement of all atoms in a protein provides the protein with its overall conformation. The highest level, quaternary structure, describes the folding of multiple protein chains that arrange into a single protein complex. Not every protein has a quaternary structure.

In proteins, allostery is the regulation of protein activity via a molecule, such as a ligand, binding to a functional site on a protein. Consequently, the allosteric hypothesis states that the effects of the binding are transmitted through the protein, often resulting in a conformational change at some other distal site [55]. However, the exact mechanisms



(a) Amino acid structure.

(b) Chain of amino acids.

Figure 1.1: The general structure of an amino acid (a) along with the structure of a linear chain of amino acids (b). The R groups represent the side-chains. Reproduced from Nature Education under a Creative Commons license [72].

behind this type of communication are still poorly understood. In particular, researchers are concerned with understanding how structural changes (if any) arise from a ligand binding event and how ligand binding affects functional activity. To address the first point, we can identify sites on the protein that are important for allosteric communication and find putative pathways within the protein that enable this communication. Linking allostery to protein function is more difficult and requires a careful analysis of protein structure.

Studying allostery is important for applications such as drug discovery and understanding disease [55]. We know that faulty allosteric processes can cause diseases, but an understanding of the functional effects of allostery can contribute towards the development of new drugs and therapeutics [69]. Studying allosteric regulation provides insights into cellular signalling and how cells respond to changes in the environment [70]. However, while there has been considerable research on the role of allostery both within a single protein and across an entire cellular signalling pathway, there is no consensus on how allostery is facilitated in even simple biological models, such as a small protein.

Allostery is often explained using a putative pathway of interacting residues. Such a pathway represents the "communication network" that an allosteric signal follows. The traditional view of allosteric signalling is that the signal is initiated from one site of the protein, usually the binding site, and propagates to a distal functional site on the protein, which is the allosteric site. Typically, the propagation of the signal induces a conformational change in the protein's three-dimensional structure, but as we will discuss in Chapter 2, this is not always the case. The structure of a protein that is not bound by any substrate, such as a ligand, is known as the apo form. When bound with a substrate, the protein structure is then called the holo form. Comparisons between these two structures are needed to understand the effects of ligand binding. The goal of allosteric models is thus to determine the pathways that signals take, identify the residues that are critical components of these pathways, and understand how allosteric processes affect overall protein function.

Researchers have devised several, oftentimes conflicting, models that attempt to explain allosteric mechanisms. Chapter 2 will explain these models in greater depth. The models we will discuss are based on computational techniques for investigating allostery. Computational techniques offer a method of analyzing complex biological systems, such as proteins, without the need to perform *in vivo* or *in vitro* experiments. As such, these methods enable researchers to study protein structures in a cost- and time-efficient manner that avoids the wet lab entirely. Different models have their own advantages and disadvantages with respect to algorithmic time complexity, representation of biological systems, and ease of use.

Our focus was on computational models that analyze protein structure fluctuations. In particular, we were interested in the interactions between side-chains. While both the backbone and the side-chains can experience fluctuations due to a ligand binding to the protein, we know that in general, side-chain atoms experience greater movement. Thus, a simple structural model of allosteric proteins would be one that has a rigid backbone (that is assumed to stay fixed between the apo and holo forms) but dynamic side-chains. Analyses for understanding allostery can then be focused on the side-chains only. Using a dynamic backbone with fixed side-chains would not make sense because it is not reasonable to assume that side-chain atoms will not move if backbone atoms are changing positions. Using a model that assumes both a dynamic backbone and dynamic side-chains would lead to a very complex analysis, and our goal was to study a simplistic model of allostery.

1.1 Our Contributions

We used a statistical analysis of side-chain motions combined with a graph-based approach to investigate allosteric mechanisms. Our method assumed that allosteric signals are propagated via side-chain interactions rather than backbone motions, and so we focused on proteins with minimal backbone deviations between the bound and unbound structures. We introduced a statistical coupling measurement called the Fluctuation Coupling Strength that measures the dynamic coupling between two interacting residues, along with a graph-based measurement called the Fluctuation Concentration that measures the overall fluctuation of a residue due to a signal propagation. The main advantage of using a graph theory method is the low computational cost to perform graph-based algorithms, enabling us to quickly run computational experiments on a variety of protein structures. Furthermore, our approach identified structural elements, such as high levels of side-chain fluctuations within an alpha helix, as well as individual residues that experience significant side-chain fluctuations. The dynamic interactions observed in these locations suggest that these structural elements and residues have a functional importance, which we argue provides greater insights into allostery beyond just a putative pathway.

1.2 Organization of this Thesis

The rest of this thesis is organized as follows. In Chapter 2, we give an overview of previous work relating to allostery and the computational techniques that have been used to study allostery. In Chapter 3, we describe our methodology, including the algorithms we employed in our study. In Chapter 4, we show the results of our computational analyses and highlight important allosteric locations identified via our algorithms. In Chapter 5, we discuss the importance of our results, the limitations of our method, and possible directions for future research.

Chapter 2

Prior Work

In this Chapter, we discuss several state-of-the-art computational allosteric communication models. Many types of computational techniques, such as Normal Mode Analysis (NMA), Elastic Network Models (ENMs), Monte Carlo (MC) and molecular dynamics (MD) simulations, Markov state models (MSMs), and graph theory methods, have been employed to determine protein regions that are important for allosteric signaling. However, the exact mechanisms that proteins employ to convey allosteric signal transmissions are still largely a mystery and are currently a hotly debated topic in the protein research community. We review the techniques researchers have used to elucidate allosteric mechanisms and we focus on identifying allosterically important residues and signaling pathways.

2.1 Models for Allostery

While allostery is largely viewed as the propagation of information from one site of a protein to a distal site, usually initiated by a ligand binding, exactly how this occurs and how this relates to the conformational changes the protein undergoes is still poorly understood. The classical view of allostery is that the ligand binding event induces a conformational change in the protein, changing it from an inactive conformation to an active conformation [94]. This concept of two dominant states, such that the ligand binding at an active site may propagate a signal to a distant binding site of the protein and change its conformation, is also referred to as the "induced fit" paradigm [1]. Since this paradigm allows allostery to be studied primarily by monitoring the individual noncovalent bonds within the protein, it has been a popular viewpoint owing to its simplicity and visual appeal [41]. However, another model of allostery does not necessitate conformational changes, and instead entropy becomes the driving force for allosteric communications [19]. With this approach, a "population shift" model was proposed, such that an ensemble of multiple states exist near the native state of a protein [93]. Ligand binding then "selects" one of these functional conformers, altering the free energy surface and leading the population to shift toward the energy of the selected conformer [58]. In addition to a ligand binding to the protein, other factors such as post-translational modifications and changes to the environment may also affect the population distribution, resulting in a more dynamic outlook for a protein's energy states [58]. In recent years, there have also been attempts to unify these two differing views using the reasoning that the existence of a communication channel between potential binding sites is what makes those sites allosteric [94].

2.1.1 Structural Allostery

In the structural view of allostery, the key principle is that structural changes produce allosteric processes in the protein. The induced fit paradigm fits within this model, as a ligand binding event would induce a structural change at the allosteric site, causing the protein to switch to its active conformation. Some researchers also argue that allosteric processes can be described via the sequential formation and breakage of individual noncovalent bonds between evolutionary conserved residues, allowing a signal to be transmitted from an effector site to a binding site [56, 21].

As evidence for the plausibility of structural allostery, pathways of bond distortions in proteins have been observed through point-mutation and thermodynamic cycle experiments [21, 81]. Furthermore, known allosteric proteins have shown substantial structural changes upon ligand binding [20, 49]. One well-known example is calmodulin, a calcium-binding messenger protein that undergoes significant backbone changes upon binding of an allosteric ligand [54]. These results indicate that analyzing the structure of a protein may provide a deeper understanding of allosteric mechanisms, and for some proteins, conformational changes between two dominant states of a protein could be the phenomenon that explains why and how intraprotein signaling occurs.

2.1.2 Dynamic Allostery

While a purely structure-centric model can be attractive due to its relative simplicity, it has been challenged since thermodynamic models show that structural changes may not be required for allostery [41]. Rather, changes in equilibrium dynamics, or the entropic component of the free energy of interactions, has been shown to play an important role in some allosteric proteins that do not undergo backbone conformational changes when a ligand binds [92]. The Gibbs free energy equation, shown in Equation 2.1, is particularly relevant here.

$$\Delta G = \Delta H - T\Delta S \tag{2.1}$$

ΔG is the change in free energy, which is dependent on the change in enthalpy ΔH, the temperature T, and the change in entropy ΔS. In a protein system, this equation measures the thermodynamic potential of interactions with other molecules, such as a ligand [48]. Some researchers suggest that the amount of conformational change an allosteric protein experiences is governed by both its enthalpic and entropic changes [92]. Specifically, proteins that are dominantly governed by enthalpy experience greater backbone structural changes, whereas proteins that are dominantly governed by entropy experience few or no backbone structural changes [92].

Some proteins can show allosteric behaviour in the absence of a structural pathway [22], as a result of surface mutations that do not induce structural changes [83], and even via disordered segments [80]. This differing viewpoint is compatible with the population shift paradigm. The basis for dynamic models of allostery comes from our long-established understanding that proteins do not exist as only a single conformation but rather in equilibrium as an ensemble of states. When the protein is not bound with a ligand, the inactive conformation substates are more "populated", meaning the equilibrium shifts towards these substates. However, when a ligand binds to the protein, the equilibrium will shift toward the active protein conformation. Experimental studies, such as with X-ray crystallography, NMR spectroscopy, and fluorescence spectroscopy, have shown side-chain fluctuations that are more consistent with this viewpoint [18]. This dynamic-centric approach offers a potentially stronger basis for allostery, as thermodynamic analyses have been developed to describe allosteric behaviour quantitatively [46, 51].

The dynamic aspects just described also give rise to an ensemble model of allostery that has become popular in recent years [41, 66]. Similar to most other dynamic views of allostery, the ensemble model emphasizes that allostery arises from energy changes within the system [49]. How much time a protein spends in a particular state of the ensemble depends on that state's stability [40], and the addition of a ligand redistributes the ensemble by changing the states' relative stabilities. Using this model, allostery can be understood in terms of the stabilities of states in the protein's native ensemble [39, 7] without needing detailed observations about structural or mechanical changes in pathways that connect the active and allosteric sites. However, this does not necessarily mean that structural changes do not play a role in allostery; rather, the ensemble view implies that structural perturbations can be reconciled with the dynamics of a protein to explain allosteric mechanisms [39].

2.2 Computational Techniques for Allostery

While experimental studies are useful for characterizing allosteric behaviour, they are usually unable to analyze entire proteins at an atomistic level, which may contain details that are crucial for understanding this behaviour [1]. In such an investigation, computational methods become especially useful. These methods encompass strategies such as statistical analyses, simulations, and modeling of residue networks, each offering predictive power that allows them to identify allosteric sites and/or pathways. While experimental methods are still an invaluable tool for investigating allostery, rapidly decreasing computational costs mean that computational techniques will only become more powerful in the future [55].

2.2.1 Normal Mode Analysis (NMA)

In normal mode analysis, the assumption is that a protein resembles an oscillating system at equilibrium, and when perturbed, some restoring force brings the system back to its equilibrium position [8]. This technique is derived from the classical mechanics description of normal (harmonic) modes, which is a pattern of motion in an oscillating system where each component of the system moves with its own consistent frequency. When applied to protein structures, the oscillating system represents the set of conformations near the minimum energy conformation, and the motions represent the dynamical changes between conformations.

NMA involves solving equations of motion, which are derived from descriptions of the potential energy and kinetic energy of the system. The goal is to compute the position of each atom at any time step subject to a small perturbation [5]. Near the equilibrium conformation $\mathbf{q}^0$, the potential energy of the system $\mathbf{q}$, $V(\mathbf{q})$, is estimated via the power series in Equation 2.2 [5]:

$$V(\mathbf{q}) = V(\mathbf{q}^0) + \sum_i \left(\frac{\partial V}{\partial q_i}\right)^0 (q_i - q_i^0) + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)^0 (q_i - q_i^0)(q_j - q_j^0) + ... \qquad (2.2)$$

Here, q_i and q_j represent the instantaneous configurations of components i and j after a perturbation, respectively. A superscript of 0 indicates equilibrium configurations. At the global energy minimum, the first term is set to zero. The second term is zero at any local minimum of the potential energy function. Thus, the second order approximation of the power series at the global energy minimum is calculated as a sum of pairwise potentials shown in Equation 2.3.

$$V(\mathbf{q}) = \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)^0 (q_i - q_i^0)(q_j - q_j^0) = \frac{1}{2}(\mathbf{q} - \mathbf{q}^0)\mathbf{H}(\mathbf{q} - \mathbf{q}^0) \tag{2.3}$$

$\mathbf{H}$ is the Hessian matrix such that each component $\mathbf{H}_{i,j}$ describes the energetic contribution of the interaction between components i and j. Since $\mathbf{H}$ is symmetric, it can be diagonalized to produce eigenvalues and eigenvectors that represent the normal modes and their respective frequencies.

However, there are some concerns associated with the validity of NMA [59]. Fluctuations are assumed to be small enough that the system practically behaves as a solid, in which atoms can vibrate with specific modes of vibration, each characterized by a specific frequency [2]. Furthermore, there is the question of whether conformational transitions are harmonic, which may be an unrealistic simplification of the protein structure's dynamics [59]. Anharmonic motions are confirmed to exist in protein structures, especially at low frequency modes, so other models may be required to reconcile these contradictory views [1]. NMA is usually performed with only C_α atoms representing each residue, thus ignoring the contribution that side-chains may offer. The assumption of harmonic fluctuations about an energetically minimized conformation indicates that this method does not take into account other motions such as local unfolding and rigid body movements [34].

Despite the limitations of NMA, it has nonetheless been used extensively for investigating allostery. Early applications of NMA on the structure of bovine pancreatic trypsin inhibitor found that the resulting modes were consistent with the directions of collective fluctuations [68, 53]. More recently, web servers for predicting allosteric communication, such as SPACER (2013) [32] and PARS (2014) [76] have been developed using NMA. The AlloPred method uses perturbations of normal modes to predict important allosteric sites on proteins [33]. Similarly, the normal modes can be calculated by observing the change in flexibility due to ligand binding to reveal allosteric sites [75].

2.2.2 Elastic Network Models (ENMs)

Similar to oscillation-driven NMA are elastic network models. In ENMs, the energy potential equations sed by NMA are replaced with a simpler harmonic potential equation, which eliminates the need for the input structure to be initially energy minimized [8]. The energy potential equation for ENMs is [6]:

$$V = \frac{\gamma}{2}\left(\sum_{ij}^{N}(R_{ij} - R_{ij}^{0})^2 f(R_{ij}^{0})\right) \tag{2.4}$$

where γ is the uniform spring constant, R_{ij}^{0} and R_{ij} are the original and instantaneous distances between residues i and j, and $f(R_{ij}^{0})$ is the Heaviside function that determines the residue pairs to be included in the summation based on an interaction cut-off distance D_c. The original distance refers to the distance obtained from the equilibrium conformation, and the instantaneous distance refers to the distance obtained from the protein structure after a perturbation. The Heaviside function is defined in Equation 2.5.

$$f(x) = \begin{cases} 1, & x < D_c \\ 0, & x \geq D_c \end{cases} \tag{2.5}$$

Different implementations of ENMs, such as the Gaussian network model (GNM) [4] and the anisotropic network model (ANM) [91], have become popular methods for studying protein dynamics. Studies have shown that the low-frequency modes identified via ENMs are similar to those found with all-atom NMA [86, 42, 10]. Consequently, ENMs can be equally as accurate in quantifying large-scale collective fluctuations as NMA methods while also requiring a lower computational cost. However, ENMs suffer many of the same drawbacks as NMA, as they are both rooted in the assumption that protein systems can be described by harmonic motions and that spring forces are acting upon the system. The usage of a uniform spring constant also draws some concerns, as it does not differentiate between different residue-residue interactions. Thus, similar to NMA, ENMs are often considered to be an oversimplification of protein dynamics.

Many coarse-grain ENMs have been implemented in web-based tools for analyzing allosteric behaviour [8]. Some of these include Hinge-Prot [28], MolMovDB [29], AD-ENM [103], and oGNM [101]. Hinge-Prot uses both GNMs and ANMs to identify hinge residues in a protein [28]. MolMovDB offers a variety of services, including a database of protein motions and a tool for predicting pathways between different conformations MolMovDB

[29]. AD-ENM is able to determine the contribution each normal mode has on a specific conformational change and can perform both coarse-grain ENM analyses and all-atom NMA AD-ENM [103]. Similarly, oGNM allows the user to use either NMA or an ENM for identifying allosteric sites [101].

2.2.3 Monte Carlo (MC) and Molecular Dynamics (MD) Simulations

While NMA and ENMs depend on various assumptions about the behaviour of the protein system, molecular dynamics simulations simply follow the movements of atoms over a time interval, using force fields based on Newtonian mechanics to simulate motion [38]. A force field refers to the functional form and parameter sets used to calculate the potential energy between interacting particles. Assuming the force field accurately reflects the mechanics of the system, MD simulations offer a more realistic representation of the dynamics of a protein compared to any other commonly used computational technique [87]. This allows for an all-atom description of protein dynamics using space and time resolutions not typically available to other methods. However, the biggest limitation of MD simulations is the large computational cost required to simulate a protein system for even relatively short (e.g. nanosecond) timescales, and allosteric processes often occur over much longer timescales [24]. Time steps in MD simulations are typically about 10 femtoseconds, and thousands of computations are usually required for each step [99].

Closely related to MD simulations are Monte Carlo simulations. However, a main difference is that MC simulations calculate thermodynamical statistical probabilities instead of employing Newtonian equations to simulate atom movements [77]. MC simulations do not provide information about timescales, but rather give the probability of changing between each conformation within the system's configuration space. As a result of the inherent randomness of MC methods, MC simulations are not deterministic, so they are used to study the behaviour of a system in thermodynamic equilibrium [77]. By applying random perturbations to the protein structure, one can obtain a sample of representative configurations under the specified thermodynamic conditions [77]. MC and MD simulations are often used in combination to take advantage of both the thermodynamic and kinetic analyses that the two techniques offer [67]. However, similar to MD simulations, a major limitation of MC simulations is the expensive computational cost.

Due to the fine-grain details that MC and MD simulations reveal, both these methods have become very popular for studying allosteric behaviour. By using perturbations to force residues into specific positions, one can observe how a distal site responds via steered

MD [44], forced MD [73], or targeted MD [71] simulations. A similar method called perturbation response scanning, which involves applying a perturbation and observing changes in neighbouring regions using MD simulations, has also been developed to identify allosteric residues [31].

2.2.4 Markov State Models (MSMs)

MD simulations provide a wealth of information about a protein's dynamic processes, but extracting such information for practical analysis is not a trivial task. Markov state models involve methods for modeling the conformational ensemble obtained from the simulation, allowing for further analyses to determine important pathways and residues for signal propagations [84]. The advantage of using MSMs in addition to MD simulations is the ability to model allosteric processes via Markov chains using only a fraction of the data needed for a full-scale simulation [84]. Instead of running a simulation over a long time period that attempts to capture the entire process from start to end, MSMs sample the ensemble of conformations that the protein undergoes over a shorter period of time, cluster the conformations into states, and compute transition probabilities between each state [12]. These transitions are then used to analyze the macroscopic behaviour of the system, and signal pathway trajectories can be generated to uncover allosteric mechanisms [43]. It should be noted that these trajectories represent transitions between conformations of the entire protein rather than pathways from residue to residue, as is typical with most descriptions of allostery [84]. Since MSMs require the usage of MD simulations to sample protein conformations, the main limitation of MSMs is the long simulation runtime required to collect enough conformations. While MSMs only need a sample of protein conformations to construct a model and do not require a full-scale simulation, this limitation prevents MSMs from performing allostery analyses as quickly as other non-MD simulation methods.

Popular software tools for constructing MSMs include PyEMMA [82] and MSMBuilder [9]. These programs automate much of the modelling process, removing the need for users to specify how the configuration space should be discretized. Another MSM approach, which uses a master equation to describe the evolution of a continuous-time Markov process rather than a typical discretized-time approach, was proposed by Long et al. in 2011 [57]. More recently, combining MD analysis packages, such as MDTraj [61] and HTMD [23], with MSM builders helps streamline the entire computational analysis pipeline and gives the MSM tools more available parameter options [43].

2.2.5 Graph-theoretic Methods

Graph theory methods represent the protein system as a network of vertices (nodes) connected by edges. As stated in Chapter 1, our model is based on a graph theory approach. These methods are rooted in the assumption that allosteric signals are propagated through a protein via atomic interactions between its residue pairs [1]. Typically, graph theory methods involve some type of residue interaction network, such that each node represents a residue in the protein and edges indicate interactions among residues. For example, the interaction may indicate that the distance between C_α atoms of two neighbouring residues is less than some threshold. Some methods choose a finer-grain representation such that each node represents an atom, and edges indicate atom-atom interactions. Another less frequently used type of residue interaction graph consists of weighted edges. One such approach is to use weights proportional to the inverse of the distance between C_β atoms of neighbouring residues [47]. Once a graph is constructed for the protein, an adjacency matrix can be derived such that each entry at row i and column j indicates the interaction between residue R_i and residue R_j. Representing the graph as a matrix then admits a variety of different graph theory methods to be applied [11].

By using a graph representation of the protein structure, these methods allow for complex analyses without the need for computationally expensive simulations. Censoni et al. found high correlation between graph centrality measures and anisotropic thermal diffusion (ATD) data, which measures the heat flow through residues [15]. In 2020, Wang et al. developed a network-based tool called OHM to identify allosteric residues and pathways by characterizing residue-residue interactions and by using a propagation algorithm [97]. Another recent study in 2021 makes use of edge-weighted residue graphs to analyze allosteric effects and protein-protein interactions [30]. In this paper, the authors use spectral decomposition to identify allosterically important residue clusters [30]. The CONTACT method, which was developed in 2013, takes as input the coordinates of a single crystal protein structure with alternative conformations and identifies putative allosteric pathways [95].

2.3 Energy Landscapes

Understanding the energy landscape is key to understanding how allostery works. Each of the computational methods outlined above makes assumptions about how a protein's energy landscape relates to allosteric processes, so an accurate model is dependent on these assumptions being realistic. According to the ensemble model, which is one of the most recognized dynamic models of allostery [66], the population distribution is correlated with the energy of each conformation in the ensemble [66]. Ligand binding changes the energy landscape, causing the stability of each conformation to shift as well [35]. Consequently, the energy landscape can be smooth or rugged, with many populated states or only a few low energy states, depending on the protein and the ligand.

One view of energy landscapes involves the idea of microstates. For this conceptualization, the protein energy surface contains many "local wells" that are separated by large energy barriers [64]. Each microstate represents a "wider well" that is the region between two barriers (see Figure 2.1) [64]. Small perturbations cause the protein to move between conformations within the same microstate, but crossing a major energy barrier into a different microstate would require larger perturbations or conformational shifts such as a rotameric change [65].

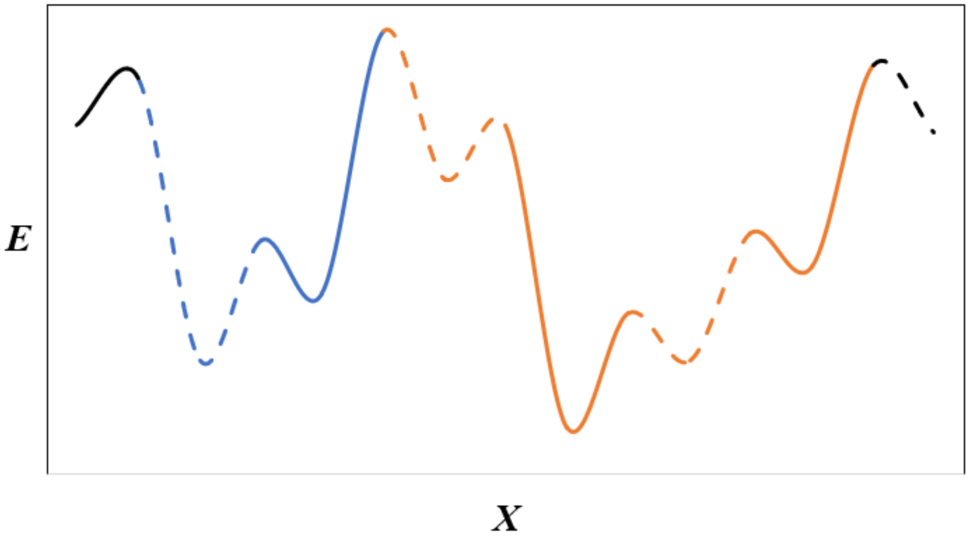


Figure 2.1: Representation of part of a protein's energy landscape, mapping the energy E as a function of the protein configuration X. Two large energy potential wells (microstates) are shown: one outlined in blue, and one outlined in orange. Each microstate consists of several smaller energy potential wells, which are differentiated by the solid and dashed lines. The microstate represented by the orange portion is likely more stable than the microstate represented by the blue portion due to a lower energy. This figure was adapted from Figure 1 of the paper by Meirovitch et al. [65]: doi:10.2174/138920309788452209.

Calculating the free energy change induced by a ligand binding to a protein could give insights on how the binding event affects the protein's dynamics [65]. For example, does the population shift into a different microstate? If so, does one part of the structure experience larger free energy changes, and could that relate to the protein's ability to propagate allosteric signals?

The role of water is also very important in determining the functional landscape of a protein [60]. In 2006, Dyson et al. found that hydrophobic forces are a driving force for protein folding [26], and contacts from water or other molecules are also frequent in a protein's native state [106]. However, many studies do not consider how the surrounding water bath affects allosteric mechanisms. Zhuravlev et al. proposed in a 2009 paper that water helps smooth the energy landscape, and that solvent degrees of freedom should be considered when evaluating effective interaction energies [104]. Thus, it is desirable for an allosteric model to accurately depict the role of water, whether through implicit solvent models or by accounting for water-mediated contacts [105].

A key contribution of this thesis is the consideration of the effects of the surrounding water bath on allosteric processes. We suggest that an allosteric signal can be initiated not just from the binding site, but also from water molecules colliding with the protein. Many studies ignore the contributions of surrounding water molecules, so our goal was to devise a more biophysically accurate model of allostery.

Chapter 3

Methods

In this Chapter we describe the methods used to create a computational model of allostery for protein systems with a fixed backbone. This allows us to focus entirely on side-chain interactions. A purely side-chain centric model provides a simpler framework for studying allostery compared to a model that needs to also consider backbone fluctuations. We view such a model as a necessary first step in understanding allostery. Without a good understanding of a simple allosteric protein system, trying to understand more complex allosteric protein systems that involve both backbone and side-chain motions would be much more difficult.

Briefly stated, the methodology involves the analysis of the correlation of side-chain motion between residue neighbours, incorporating information about the degrees of freedom of atom-atom contacts and the directionality of allosteric signal propagation. The overall steps of our approach are as follows:

1. Add hydrogen atoms and perform an initial energy minimization of the protein structure
2. Derive a neighbourhood list for each residue
3. Investigate the degrees of freedom of atoms involved in residue-residue interactions
4. Perform perturbations followed by energy minimization for each residue
5. Analyze side-chain motion correlations
6. Evaluate fluctuation propagations

7. Identify high fluctuation sites from the fluctuation data and residue networks

The energy minimizations and visualization were performed using the UCSF Chimera software [79], which is an interactive molecular visualization program developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco. Chimera supports Python scripting via its IDLE interactive development environment and statements in a Python script can be used to retrieve files containing atomic coordinates, e.g., PDB format files, for 3D visualization and subsequent processing.

3.1 Initial Energy Minimization

The first step in our computational pipeline is to preprocess the protein conformational data. Using the protein's conformation from its PDB file, hydrogen atoms were placed in the structure using the "addh" command in Chimera if they are not already included in the PDB file. Then, we used Chimera's Minimize Structure tool, which performs energy minimization on the molecular structure. This process derives a conformation that is consistent with a local energy well within the current microstate, using Amber force fields [98]. While the resultant structure was not guaranteed to have the global minimum energy, the procedure helped the system move towards a local energy minimum without crossing major energy barriers, meaning the system was likely to stay within the same microstate when perturbed. The initial energy minimization eliminated the crystal packing effect that may be present in the PDB conformation [45]. The addition of hydrogen atoms helped in creating more realistic energy minimized structures for both the initial minimization and all subsequent energy minimizations.

Minimize structure works by first performing a steepest descent algorithm to prune conformations that result in highly unfavourable clashes, followed by a conjugate gradient minimization algorithm to find a local energy minimum. The conjugate gradient minimization was slower than the steepest descent minimization but was more effective in searching for energy minima after unfavourable clashes have been removed from the search space. By default, the parameter settings used for steepest descent minimization were 100 steps and a stepsize of 0.02Å, and conjugate gradient minimization used 10 steps with a stepsize of 0.02Å. We experimented with several different parameter values to test the trade-off between speed and accuracy, and determined that the default parameter values were the best choices to use. This was done by first performing a baseline minimization using a large number of steps for minimization to obtain a conformation with a relatively low energy, and then performing more minimizations using a variety of parameter values. After each run, every residue's dihedral angles and the time spent for minimization were recorded and compared with the baseline values. We found that on average, minimizations with the default parameters were the fastest while still able to yield very accurate conformations (side-chain dihedral angles within 1° on average relative to the baseline values).

3.2 Deriving Neighbour Lists

Next, we generated a list of all neighbouring residues for each residue within the protein. We used a method for identifying pairs of neighbouring residues similar to that described by Cohen et al. [17]. They used four distances between a pair of residues to describe their interaction, which differs from the traditional method of simply measuring the distance between C_α atoms of two different residues to determine possible interaction. They showed that this method has more information content due to the asymmetry of many residue-residue interactions, thus yielding a more precise measure of side-chain interactions. Thus, the four-distances measure provides a better approximation of forces between any two residues and can more accurately describe residue proximity.

The four-distances strategy works as follows: first, for each residue R_i in the protein, nearby candidate neighbours were selected if their C_β atoms were within some predefined distance D_1 from the C_β atom in R_i. Then, we measured the pair-wise distances between two atoms from R_i and two atoms from R_j. We used a table, shown in Table A.1, derived from Cohen et al.'s data [17] to determine the atoms that are involved in the calculations. Let the positions of the two atoms from R_i be R_i^a and R_i^b, and let the positions of the two atoms from R_j be R_j^a and R_j^b. To filter candidate neighbour residues, we determined if $min(|R_i^a - R_j^a|, |R_i^a - R_j^b|, |R_i^b - R_j^a|, |R_i^b - R_j^b|) < D_2$, where D_2 is some distance threshold that is less than D_1. If the inequality held, R_j was selected as a neighbour of R_i

For example, if R_i is an arginine residue and R_j is a glutamic acid residue, the two atoms from R_i would be CD and NH2, and the two atoms from R_j would be OE1 and OE2 (see Table A.1). The four distances used would be the distances from CD to OE1, CD to OE2, NH2 to OE1, and NH2 to OE2. If at least one of the distances was less than D_2, then that residue was classified as a neighbour of R_i. Figure 3.1 shows a visual display of the distance measures for this example.

D_1 was set to 10Å, and D_2 was set to 5Å. Alanine and glycine residues along with cysteine residues involved in disulphide bridges were excluded in this step and thus were not part of the residue interaction graph. Alanine and glycine do not have side-chain motions due to a lack of χ angles. Similarly, cysteine residues involved in disulphide bridging have

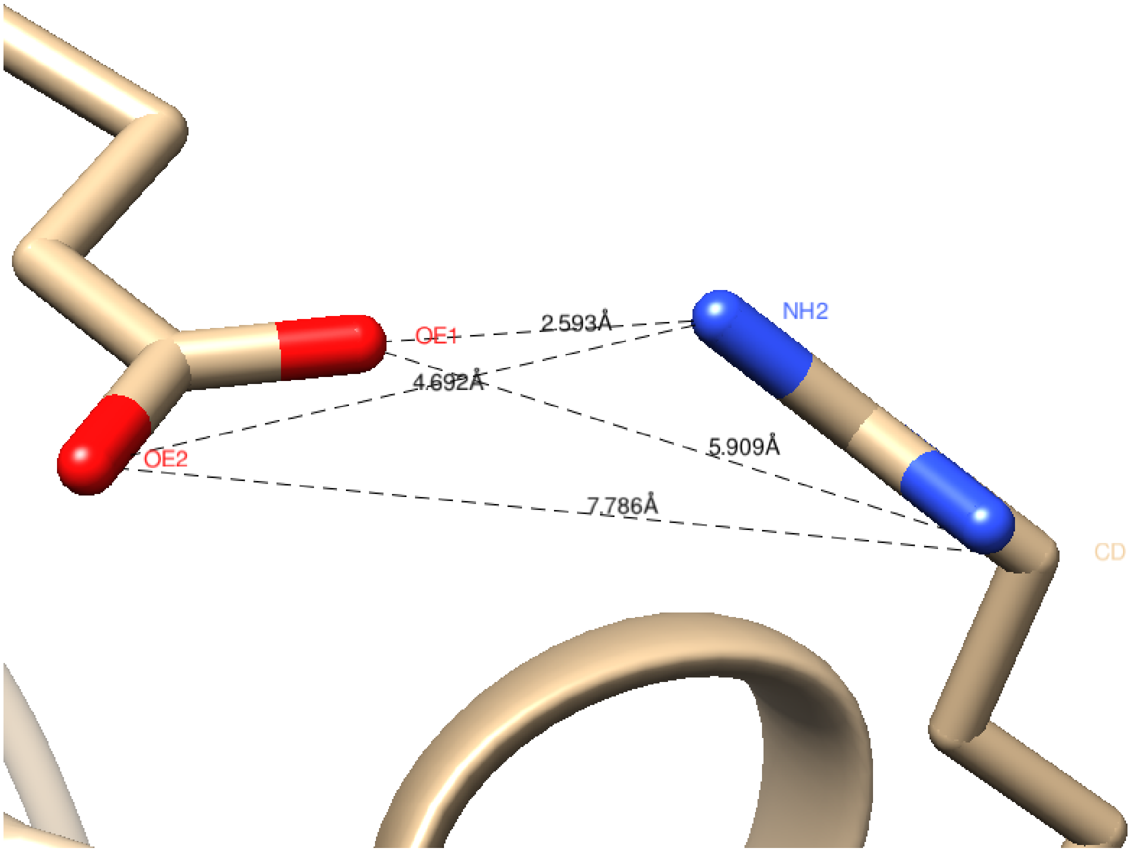


Figure 3.1: Display showing the four distances used to determine the interaction between an arginine residue (right) and a glutamic acid residue (left). The two atoms used from the arginine residue to calculate the distances were CD and NH2, while the two atoms used from the glutamic acid residue were OE1 and OE2. The dashed black lines represent the atom-atom distance measures that were calculated.

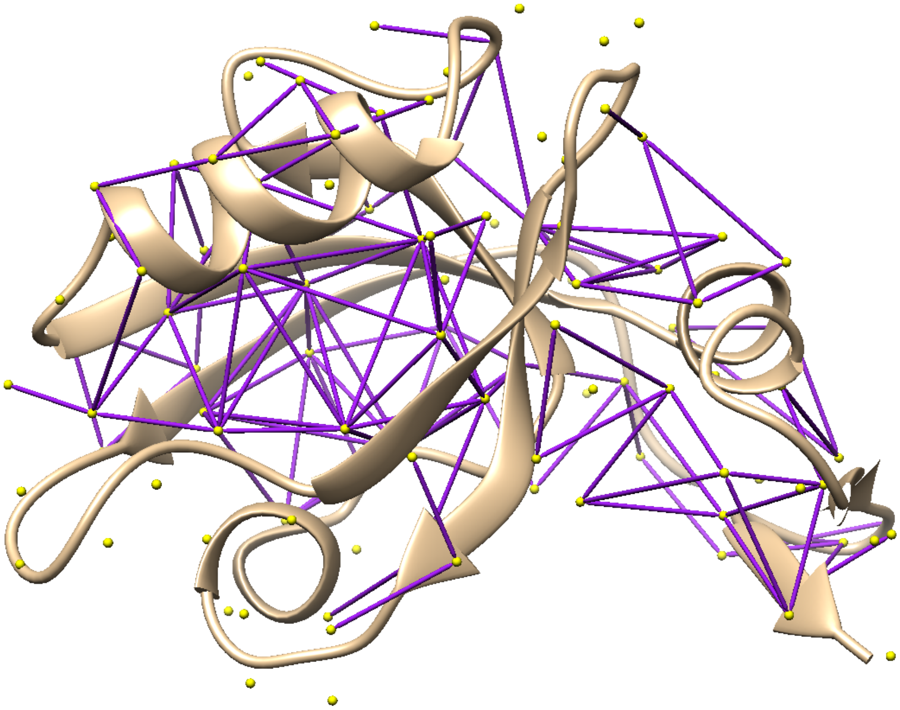


Figure 3.2: A residue interaction network of the PDZ3 domain of the PSD-95 protein (PDB ID: 1BFE). Nodes are represented as yellow spheres at the centroid of each residue and edges are represented by purple spindles connecting two nodes. This display was created with the help of the Python StructBio package for Chimera [14].

very little, if any, side-chain motions due to the rigidity of the linkage. Consequently, it was not necessary to perform side-chain correlation analyses involving these residues.

We used residue neighbour lists to derive a residue interaction network representing the protein. Figure 3.2 shows an example of such a network. This network is a representation of the protein structure such that nodes represent residues and an edge between two nodes indicates that the two connected residues are interacting. A modified version of the basic residue interaction network will later be a key component of the propagation analysis described in Section 3.6.2.

3.3 Degrees of Freedom and Atom-Atom Contacts

While not necessary for the side-chain fluctuation analysis that our model is based on, we wanted to derive more information about the nature of side-chain interactions. Specifically, we were interested in analyzing how the degrees of freedom (DoF) of side-chain atoms affected side-chain fluctuations. The DoF refers to the number of independent motions that are allowed on a physical body. In the context of protein structures, the DoF of a side-chain atom is determined by the number of residue χ angles that can affect the position of said atom. Assuming a fixed backbone and fixed bond lengths, the position of each side-chain atom of a residue can be precisely determined via the residue's χ angles. We performed a statistical analysis to investigate the distribution of different DoF of interactions between neighbouring residues. Here, we define the DoF of an interaction as the pair of DoF values that represent the DoF of the closest pair of side-chain atoms between the two interaction residues. Let the closest pair of atoms between two neighbours residues R_i and R_j be a_i (from R_i) and a_j (from R_j). The DoF of the interaction is $DoF_i : DoF_j$, such that DoF_i is the DoF of a_i and DoF_j is the DoF of a_j.

Atoms with a lower DoF have a lower range of possible motions compared to atoms with a higher DoF. We used a collection of high resolution protein data taken from the Dunbrack lab PISCES database [96] and compiled side-chain interactions within the proteins into a side-chain atlas. From this atlas, we performed an extensive sampling of residue-residue interactions and evaluated the closest pair of atoms between interacting residues. The interactions were separated into categories based on their DoF and their secondary structure membership (alpha helix, beta sheet, or strand). Table 3.1 shows the results of this analysis. Among all pair-wise DoF interactions, DoF1:DoF1 interactions are the most common, followed by DoF1:DoF2 interactions.

3.4 Perturbations and Energy Minimizations

To evaluate correlated motions between side-chains, we needed to obtain a set of side-chain conformations that represented a limited range of motions exhibited by the residue pair without changing rotameric settings. This was done once all pairs of neighbouring residues have been identified. To simulate side-chain motions, we applied a series of perturbations to side-chains followed by energy minimizations.

For a residue R_a and a neighbouring residue R_b, we perturbed R_a, performed energy minimization on the protein structure, and recorded the resulting side-chain dihedral angles

	Helix-Helix	Helix-Loop	Helix-Strand	Loop-Loop	Loop-Strand	Strand-Strand	Total	% of All
1-1	83920	40525	7019	75502	25576	26192	258734	36.55%
1-2	71995	47434	8733	65932	24258	21863	240215	33.93%
2-2	29848	20523	3770	23942	9344	7705	95132	13.44%
1-3	12031	8137	1499	7941	3636	2885	36129	5.10%
2-3	11824	7610	1354	6988	3355	2691	33822	4.78%
2-5	3997	3302	490	2872	1245	809	12715	1.80%
1-5	3719	3119	431	3031	1145	801	12246	1.73%
2-4	2146	1484	234	1478	534	365	6241	0.88%
1-4	1233	1291	201	1247	463	294	4729	0.67%
3-3	1245	633	105	556	320	314	3173	0.45%
3-5	990	604	75	491	264	196	2620	0.37%
3-4	429	277	54	210	118	60	1148	0.16%
5-5	133	108	15	142	58	22	478	0.07%
4-5	136	108	20	76	66	20	426	0.06%
4-4	44	19	3	22	15	11	114	0.02%

Table 3.1: Frequency of DoF interactions based off the closest pair of side-chain atoms between two interacting residues, using a sampling of 707922 interacting residue pairs from various high-resolution proteins in the PDB. The column headers indicate the secondary structure the two residues are part of, and the row headers indicate the degrees of freedom of the two interacting atoms.

of R_a and R_b. We then repeated this step with other perturbations of R_a to obtain a set of conformations that represent a range of side-chain motions for R_a and R_b. The perturbations involved changing a residue's χ_1 angle and/or its χ_2 angle, if applicable, to settings $\chi \in \{q_0 + ds | d \in \{-4, -3, ..., 3, 4\}\}$. Here, q_0 is the initial χ angle setting and s is the stepsize that we set to 10°. For a residue with only a χ_1 side-chain dihedral angle, there were a total of 9 perturbations. A χ_2 perturbation is applied for each χ_1 perturbation, so for any residue that has both a χ_1 and a χ_2 angle, there were a total of $9 \times 9 = 81$ final conformations generated. Note that depending on the residue types, the number of conformations generated for the residue pair R_a and R_b may not have been the same if R_b was instead the perturbed residue with R_a acting as its neighbour. For example, if R_a was the perturbed residue and has χ_1 and χ_2 angles, and R_b was a neighbour of R_a and has only a χ_1 angle, there would be 81 generated conformations. However, if R_b was the perturbed residue, there would be only 9 generated conformations.

Energy minimizations also implicitly take into consideration the effect of a perturbation on all the neighbours of the perturbed residue, rather than just a single residue and one of its neighbours. Furthermore, since energy minimizations are performed with respect to the entire protein system, each minimization also considers the effects of further away residues that may not be immediate neighbours of a residue when deciding that residue's conformation. Thus, this procedure accounts for both local changes and global changes within a protein structure.

We decided to restrict perturbations to χ_1 and χ_2 angles, even for residues that had more than two χ angles. This was because the majority of residues with three or more χ angles were found outside the hydrophobic core of proteins [100], indicating that allosteric signal transmission were likely to be conveyed by residues with one or two χ angles. Limiting the number of minimizations needed for each pair of interacting residues facilitated a shorter computational runtime, as energy minimizations in Chimera have a heavy computational cost.

3.5 Side-chain Correlation Analysis

After we generated a conformational sampling for each interacting residue pair, we examined the side-chain motion exhibited in the conformations. We used canonical correlation analysis (CCA) to measure the correlation between two sets of multivariate data, and we used a kernel function to express the conformational variables in terms of side-chain dihedral angles. When put together, this method is known as a kernelized canonical correlation analysis (KCCA) [88]. The purpose of this method was to derive a measure that approximates the degree to which side-chain interactions are coupled.

3.5.1 Canonical Correlation Analysis

For two univariate random variables $\mathbf{X}$ and $\mathbf{Y}$, their correlation can simply be computed via their Pearson correlation coefficient. However, if $\mathbf{X}$ and $\mathbf{Y}$ are multivariate random variables of a observations with $\mathbf{X} \in \mathbb{R}^{a \times d_x}$ and $\mathbf{Y} \in \mathbb{R}^{a \times d_y}$, the evaluation of correlation becomes more complicated as a simple Pearson correlation computation cannot be used [37]. To handle multivariate data, we needed another method of correlation analysis. One measure that can be used to compute correlations with multivariate data is the canonical correlation coefficient. The canonical correlation between two multivariate variables $\mathbf{x} \in \mathbb{R}^{d_x}$ and $\mathbf{y} \in \mathbb{R}^{d_y}$ is

$$\rho(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{u} \in \mathbb{R}^{d_x}, \mathbf{v} \in \mathbb{R}^{d_y}} Corr(\mathbf{u}^\top \mathbf{x}, \mathbf{v}^\top \mathbf{y}) \tag{3.1}$$

with

$$Corr(\mathbf{u}^\top \mathbf{x}, \mathbf{v}^\top \mathbf{y}) = \frac{\mathbf{u}^\top Cov(\mathbf{x}, \mathbf{y})\mathbf{v}}{\sqrt{(\mathbf{u}^\top Var(\mathbf{x})\mathbf{u})(\mathbf{v}^\top (Var(\mathbf{y})\mathbf{v})}} \tag{3.2}$$

and Cov and Var representing the covariance and variance functions, respectively [37]. The goal is to find weight vectors $\mathbf{u}$ and $\mathbf{v}$ that will maximize the correlation between the two variables $\mathbf{x}$ and $\mathbf{y}$. Note that the right hand side of Equation 3.2 can be rewritten as

$$\max_{\mathbf{u}\in\mathbb{R}^m, \mathbf{v}\in\mathbb{R}^n} \mathbf{u}^\top [Cov(\mathbf{x}, \mathbf{y})]\mathbf{v} \tag{3.3}$$

subject to the constraints

$$\mathbf{u}^\top [Var(\mathbf{x})]\mathbf{u} = 1 \quad \text{and} \quad \mathbf{v}^\top [Var(\mathbf{y})]\mathbf{v} = 1. \tag{3.4}$$

With a dataset of observations $\mathbf{X}$ and $\mathbf{Y}$, the Equation 3.1 becomes

$$\hat{\rho}(\mathbf{X}, \mathbf{Y}) = \max_{\substack{\mathbf{u}^\top \mathbf{X}^\top \mathbf{X}\mathbf{u}=1 \\ \mathbf{v}^\top \mathbf{Y}^\top \mathbf{Y}\mathbf{v}=1}} \mathbf{u}^\top \mathbf{X}^\top \mathbf{Y}\mathbf{v}. \tag{3.5}$$

3.5.2 A Modified von Mises Kernel Function

Given two interacting residues R_x and R_y, the conformational sample consisted of a set of a dihedral angles observations $\mathbf{X} \in \mathbb{R}^{a\times d_x}$ for R_x and a set of a dihedral angle observations $\mathbf{Y} \in \mathbb{R}^{a\times d_y}$ for R_y, in which $d_x \in \{1, 2, 3, 4\}$ represented the number of dihedral angles in R_x and $d_y \in \{1, 2, 3, 4\}$ represented the number of dihedral angles in R_y. Since dihedral angles assume angular values in the range[-180, 180] and d_x may not equal d_y, a suitable kernel function was needed to represent similarity.

Kernel functions operate in a high-dimensional, implicit feature space without explicitly performing computations on the high-dimensional data. The kernel function calculates a "feature mapping" $\varphi : X \to V$ of all pairs of the lower-dimensional input data in the input space X to the higher-dimensional implicit feature space V such that the kernel function can be expressed as an inner product in V. For a pair of data points $\mathbf{x}$ and $\mathbf{x}'$ from the input data, the kernel mapping satisfies $K(\mathbf{x}, \mathbf{x}') = \langle\varphi(\mathbf{x}), \varphi(\mathbf{x}')\rangle_V$. As long as V is an inner product space, we do not need an explicit representation for φ.

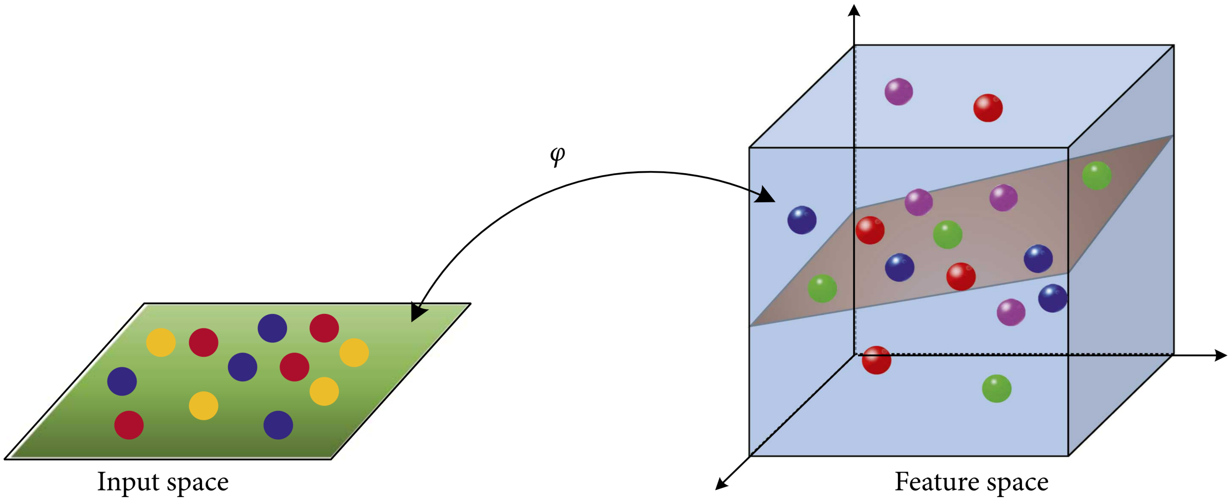


Figure 3.3: Example of a kernel mapping from a lower-dimension input space to a higher-dimension feature space. Image reproduced from Figure 3 of the paper by Mei and Tan [63]: https://doi.org/10.1155/2021/5583389.

In our case, we wanted a kernel function that could transform the angular data of the input space to a higher-dimensional feature space for easier comparison. We chose to use a modified von Mises kernel function [88]. For a pair of vector dihedral angle observations $\mathbf{x_i}, \mathbf{x_j} \in \mathbb{R}^{d_x}$ of a residue R_x, the kernel function calculates

$$K(\mathbf{x_i}, \mathbf{x_j}) = \frac{e^p}{(\prod_{m=1}^{d_x} I_0(m))^{d_x}} \tag{3.6}$$

where

$$p = \sum_{m=1}^{d_x} (\kappa_m \prod_{t=1}^{m} \cos(x_{it} - x_{jt})). \tag{3.7}$$

I_0 is the modified Bessel function of order 0 [16] and κ_m is a custom kernel parameter that was set according to the type and rotameric setting of R. Variables x_{it} and x_{jt} represent scalar components of the vectors $\mathbf{x_i}$ and $\mathbf{x_j}$, respectively. The intuition behind this modified von Mises kernel function was that the position of an atom with a higher degree of freedom was dependent on all lower degree dihedral angles, and so dihedral angles should not have been treated as independent variables but rather as related to all their lower degree of freedom dihedral angles as well. To give an example, the position of the CG atom on an arginine residue is affected by the residue's χ_1 and χ_2 angles, assuming a fixed backbone. Changes to only χ_2 provide one degree of freedom to the CG atom if χ_1 is fixed; however, when χ_1 is also changing, the position of the CG atom has two degrees of freedom. Thus, in this case, the effects of χ_2 on the position of the CG atom is dependent not only on its own value but also the value of χ_1. The modified Bessel function was used to get a better kernel density estimate and to reduce the difference between kernel values obtained with different values of d_x [90].

As previously stated, the kernel concentration parameter κ_m has a unique value for each residue type and rotameric setting. This value was obtained by extracting the standard deviation of χ_m considering all residues of the same type and rotameric setting in the Dunbrack backbone-dependent rotamer library [85]. Therefore, the value of κ_m was correlated with the standard deviation of each dihedral angle for a particular residue, based on extensive sampling of actual protein data.

If we reparameterize weight vectors $\mathbf{u}$ and $\mathbf{v}$ as $\mathbf{u} = \mathbf{X}^\top \boldsymbol{\alpha}$ and $\mathbf{v} = \mathbf{Y}^\top \boldsymbol{\beta}$, Equation 3.5 becomes

$$\hat{\rho}(\mathbf{X}, \mathbf{Y}) = \max_{\substack{\boldsymbol{\alpha}^\top \mathbf{X}\mathbf{X}^\top \mathbf{X}\mathbf{X}^\top \boldsymbol{\alpha}=1 \\ \boldsymbol{\beta}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{Y}\mathbf{Y}^\top \boldsymbol{\beta}=1}} \boldsymbol{\alpha}^\top \mathbf{X}\mathbf{X}^\top \mathbf{Y}\mathbf{Y}^\top \boldsymbol{\beta}. \tag{3.8}$$

Let $\mathbf{K_x}$ and $\mathbf{K_y}$ denote the kernelized matrices of the dihedral angle datasets $\mathbf{X}$ and $\mathbf{Y}$, respectively, such that $\mathbf{K_x} = \mathbf{X}\mathbf{X}^\top$ and $\mathbf{K_y} = \mathbf{Y}\mathbf{Y}^\top$. The kernelized canonical correlation analysis (KCCA) problem can now be formulated as finding $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ such that

$$\hat{\rho}(\mathbf{X}, \mathbf{Y}) = \max_{\substack{\boldsymbol{\alpha}^\top \mathbf{K_x}^2 \boldsymbol{\alpha}=1 \\ \boldsymbol{\beta}^\top \mathbf{K_y}^2 \boldsymbol{\beta}=1}} \boldsymbol{\alpha}^\top \mathbf{K_x}\mathbf{K_y}\boldsymbol{\beta}. \tag{3.9}$$

The corresponding Lagrangian of Equation 3.9 is

$$L(\lambda, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\alpha}^\top \mathbf{K_x}\mathbf{K_y}\boldsymbol{\beta} - \frac{\lambda}{2}(\boldsymbol{\alpha}^\top \mathbf{K_x}^2 \boldsymbol{\alpha} - 1) - \frac{\lambda}{2}(\boldsymbol{\beta}^\top \mathbf{K_y}^2 \boldsymbol{\beta} - 1). \tag{3.10}$$

The derivatives of the Lagrangian with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are

$$\mathbf{K_x}\mathbf{K_y}\boldsymbol{\beta} - \lambda \mathbf{K_x}^2 \boldsymbol{\alpha} = 0 \tag{3.11}$$

$$\mathbf{K_y}\mathbf{K_x}\boldsymbol{\alpha} - \lambda \mathbf{K_y}^2 \boldsymbol{\beta} = 0. \tag{3.12}$$

Thus, the maximization problem in Equation 3.9 can be reformulated as

$$\boldsymbol{I\alpha} = \lambda^2 \boldsymbol{\alpha} \tag{3.13}$$

which is a standard eigenvalue problem of the form $\mathbf{Ax} = \lambda \mathbf{x}$, such that λ represents the eigenvalues.

3.5.3 Incomplete Cholesky Decomposition and Regularization

As the size of the data grows, the sizes of the kernel matrices grow at a quadratic rate, since for a dataset $\mathbf{X}$ of a observations, the kernel matrix $\mathbf{K_x}$ will be of size $a \times a$. To reduce the complexity of operations on the kernel matrices, we can perform a decomposition on the kernel matrices, such as the Cholesky decomposition. However, for larger datasets, even the calculation of the Cholesky decomposition of the kernel matrices can be computationally expensive [37]. Furthermore, we can obtain maximal correlation if the kernel matrices are invertible, giving a trivial solution. To counteract these problems, incomplete Cholesky decomposition was applied to reduce the dimensionality of $\mathbf{K_x}$ and $\mathbf{K_y}$, and regularization was used to avoid a trivial solution [37].

The incomplete Cholesky decomposition (ICD) of a symmetric, positive definite matrix $\mathbf{M}$ is $\mathbf{M} \approx \mathbf{LL}^\top$, where $\mathbf{L}$ is a lower triangular matrix. This is essentially a less computationally expensive approximation of the Cholesky decomposition of $\mathbf{M}$. ICD is often used to solve the KCCA problem, as it is useful for computing the required eigenstructure [3]. Our ICD algorithm, shown in Algorithm 1, is a modified version of the algorithm presented by Harbrecht et al. [36].

Algorithm 1 Incomplete Cholesky Decomposition

```
    Input symmetric positive definite $N \times N$ matrix $\mathbf{A}$ and error tolerance $\epsilon > 0$
    Output lower triangular matrix $\mathbf{L}$ such that $trace(\mathbf{A} - \mathbf{L}\mathbf{L}^\top) \leq \epsilon$
procedure ICD
    $i \leftarrow 1$
    For $j \in [1, N]$, $\mathbf{L}_{j,j} \leftarrow \mathbf{A}_{j,j}$
    $\mathbf{p} \leftarrow (1, 2, ..., n)$
    $\mathbf{d} \leftarrow diag(\mathbf{L})$
    $error \leftarrow \|\mathbf{d}\|$
    while $error > \epsilon$ and $i \leq N$ do
        if $i > 0$ then        ▷ Find new best element $j^*$ and apply permutation
            $j^* \leftarrow argmax(\mathbf{d}) + i$
            swap $\mathbf{p}_i$ and $\mathbf{p}_{j^*}$
            swap $\mathbf{L}_{i,1:i-1}$ and $\mathbf{L}_{j^*,1:i-1}$
        else        ▷ First iteration, set best element $j^*$ as 0
            $j^* \leftarrow 0$
        $\mathbf{L}_{i,i} \leftarrow \sqrt{\mathbf{p}_{j^*}}$        ▷ Set diagonal element $\mathbf{L}_{i,i}$
        $\mathbf{L}_{i+1:N,i} \leftarrow \frac{1}{\mathbf{L}_{i,i}}(\mathbf{A}_{i+1:N,i} - \sum_{j=1}^{i-1} \mathbf{L}_{i+1:N,j}\mathbf{L}_{i,j})$        ▷ Calculate column $i$ of $\mathbf{L}$
        For $j \in [i+1, N]$, $\mathbf{L}_{j,j} \leftarrow \mathbf{A}_{j,j} - \sum_{k=1}^{i} \mathbf{L}_{j,k}^2$        ▷ Update diagonal elements of $\mathbf{L}$
        $\mathbf{d} \leftarrow \sum_{k=j}^{n} diag(\mathbf{L})_k$
        $error \leftarrow \|\mathbf{d}\|$
        $i \leftarrow i + 1$
```

Regularizations are techniques used to avoid overfitting on the training set, which in our case could result in maximal correlations if $\mathbf{K_x}$ and $\mathbf{K_y}$ were invertible. We performed regularization by adding a small weight τ to the constraints such that the optimization problem in Equation 3.9 was subject to

$$\boldsymbol{\alpha}^\top {\mathbf{K_x}}^2 \boldsymbol{\alpha} + \tau \boldsymbol{\alpha}^\top \mathbf{K_x} \boldsymbol{\alpha} = 1 \tag{3.14}$$

$$\boldsymbol{\beta}^\top {\mathbf{K_y}}^2 \boldsymbol{\beta} + \tau \boldsymbol{\beta}^\top \mathbf{K_y} \boldsymbol{\beta} = 1. \tag{3.15}$$

3.5.4 The Kernelized Canonical Correlation Algorithm

We will now describe the complete kernelized canonical correlation analysis algorithm. From the data matrices $\mathbf{X}$ and $\mathbf{Y}$ that represented the dihedral angle observations for a perturbed residue and its neighbouring residue, respectively, we created the kernel matrices $\mathbf{K_x}$ and $\mathbf{K_y}$. Regularization was applied to the diagonals of the kernel matrices, which were then decomposed via ICD to produce the lower triangular matrices $\mathbf{L_x}$ and $\mathbf{L_y}$.

$$\begin{aligned}\mathbf{K_x} &\approx \mathbf{L_x}\mathbf{L_x}^\top \\ \mathbf{K_y} &\approx \mathbf{L_y}\mathbf{L_y}^\top.\end{aligned}$$

Substituting the new representations into Equations 3.11 and 3.12 and multiplying the first equation by $\mathbf{L_x}^\top$ and the second equation by $\mathbf{L_y}^\top$ gives

$$\mathbf{L_x}\mathbf{L_x}^\top\mathbf{L_y}\mathbf{L_y}^\top\boldsymbol{\beta} - \lambda\mathbf{L_x}\mathbf{L_x}^\top\mathbf{L_x}\mathbf{L_x}^\top\boldsymbol{\alpha} = 0 \tag{3.16}$$

$$\mathbf{L_y}\mathbf{L_y}^\top\mathbf{L_x}\mathbf{L_x}^\top\boldsymbol{\alpha} - \lambda\mathbf{L_y}\mathbf{L_y}^\top\mathbf{L_y}\mathbf{L_y}^\top\boldsymbol{\beta} = 0. \tag{3.17}$$

We used ICD to approximate the kernel matrices, allowing us to re-represent the correlations with reduced dimensionality. We accomplished this by generating the $\mathbf{Z}$ matrices, which represent the new correlation matrices.

$$\mathbf{Z_{xx}} = \mathbf{L_x}^\top\mathbf{L_x} \tag{3.18}$$

$$\mathbf{Z_{yy}} = \mathbf{L_y}^\top\mathbf{L_y} \tag{3.19}$$

$$\mathbf{Z_{xy}} = \mathbf{L_x}^\top\mathbf{L_y} \tag{3.20}$$

$$\mathbf{Z_{yx}} = \mathbf{L_y}^\top\mathbf{L_x}. \tag{3.21}$$

Let $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ represent the new weight vectors with reduced dimensionality, such that

$$\tilde{\boldsymbol{\alpha}} = \mathbf{L_x}^\top\boldsymbol{\alpha} \tag{3.22}$$

$$\tilde{\boldsymbol{\beta}} = \mathbf{L_y}^\top\boldsymbol{\beta} \tag{3.23}$$

Substituting into Equations 3.16 and 3.17 and multiplying the first equation by $\mathbf{Z_{xx}}^{-1}$ and the second equation by $\mathbf{Z_{yy}}^{-1}$ gives

$$\mathbf{Z_{xy}}\tilde{\boldsymbol{\beta}} - \lambda\mathbf{Z_{xx}}\tilde{\boldsymbol{\alpha}} \tag{3.24}$$

$$\mathbf{Z_{yx}}\tilde{\boldsymbol{\alpha}} - \lambda\mathbf{Z_{yy}}\tilde{\boldsymbol{\beta}}. \tag{3.25}$$

We can rewrite $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ as

$$\tilde{\boldsymbol{\beta}} = \frac{\mathbf{Z_{yy}}^{-1}\mathbf{Z_{yx}}\tilde{\boldsymbol{\alpha}}}{\lambda} \tag{3.26}$$

$$\mathbf{Z_{xy}}\mathbf{Z_{yy}}^{-1}\mathbf{Z_{yx}}\tilde{\boldsymbol{\alpha}} = \lambda^2\mathbf{Z_{xx}}\tilde{\boldsymbol{\alpha}}. \tag{3.27}$$

Let $\mathbf{SS}^\top$ be the complete Cholesky decomposition of $\mathbf{Z_{xx}}$ with regularization applied; that is, $\mathbf{SS}^\top = \mathit{cholesky}((1-\tau)\mathbf{Z_{xx}}+\tau\mathbf{I})$. Let $\hat{\boldsymbol{\alpha}} = \mathbf{S}^\top\tilde{\boldsymbol{\alpha}}$. We then setup the new eigenvalue problem defined as:

$$\mathbf{S}^{-1}((1-\tau)\mathbf{Z_{xy}} + \tau\mathbf{I})\mathbf{Z_{yx}}(\mathbf{S}^{-1})^\top\hat{\boldsymbol{\alpha}} = \lambda^2\hat{\boldsymbol{\alpha}} \tag{3.28}$$

This is a symmetric eigenvalue problem of the form $\mathbf{Ax} = \lambda\mathbf{x}$ and it can be solved using any standard eigenvalue problem solver. We used the Python NumPy library's numpy.linalg.eig function. The canonical correlation weight $\tilde{\boldsymbol{\alpha}}$ can be computed as $(\mathbf{S}^{-1})^\top\hat{\boldsymbol{\alpha}}$, and $\tilde{\boldsymbol{\beta}}$ can be computed via Equation 3.26. We will refer to the final correlation value as the Fluctuation Coupling Strength (FCS), with a value in the range [0, 1].

3.6 Propagation Analysis

Following the correlation analysis, we created an edge-weighted digraph representing a residue network for the protein. Similar to the basic residue interaction network mentioned in Section 3.2, each node in the graph represents a residue. However, edges are now directed and weighted, with an edge (i, j) representing a fluctuation transmission from residue R_i to another residue R_j, and the weight $w(i, j)$ representing the FCS from R_i to R_j. A display of this modified residue interaction graph, which we will call the residue correlation network, is shown in Figure 3.4.

The FCS values were combined into a correlation matrix $\mathbf{A}$. Each entry $\mathbf{A_{ij}}$ was the FCS from residue R_i to a neighbouring residue R_j. Due to the non-symmetric nature of our energy minimization conformation generating method, $\mathbf{A}$ was not a symmetric matrix. This FCS matrix acted as the basis for the algorithms presented in Sections 3.6.1 and 3.6.2.

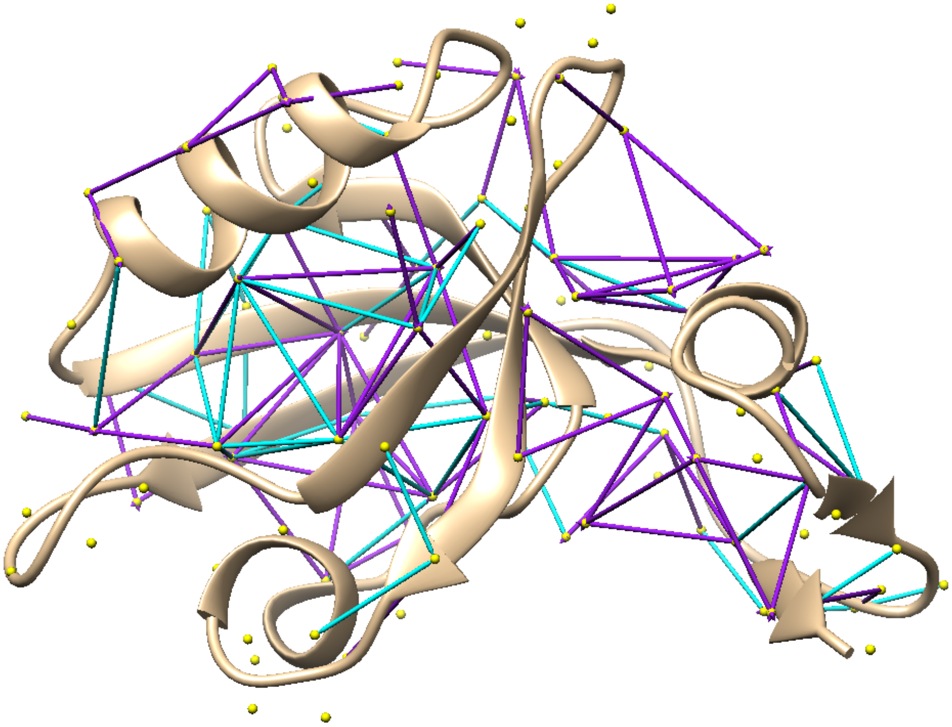


Figure 3.4: A residue correlation network of the PDZ3 domain of the PSD-95 protein (PDB ID: 1BFE). Nodes are represented as yellow spheres at the centroid of each residue. Edges between two interacting residues are represented with either a single purple arrow or a cyan spindle. A purple arrow going from a residue R_i to a neighbouring residue R_j indicates that the FCS from R_i to R_j is at least 1.2 times higher than the FCS from R_j to R_i. A cyan spindle indicates that the FCS values between the two residues are within a factor of 1.2.

3.6.1 ResidueRank

ResidueRank was the first algorithm we devised for revealing important allosteric residues. This algorithm was named after the initial PageRank algorithm that was used by Google to rank webpages [74]. The premise of our algorithm is similar, but applied to residues within a protein rather than webpages on the Internet. Residues were ranked based on the "links" (correlations) they have with other residues, with the assumption that more important residues would have highly correlated side-chain motions with respect to its neighbours. As with PageRank, the ResidueRank weight of a residue was determined recursively, as it took into consideration both the correlations it had with its neighbours and the ResidueRank of those neighbours.

In the original PageRank algorithm, the output was a probability distribution that represented how likely it was that a person would arrive at a particular webpage by randomly clicking on links. With ResidueRank, the output could be interpreted as a numerical weighting of residues describing the relative importance of residues in allosteric signal transmissions. Similar to PageRank, the sum of all the residue weights would equal 1.

The basis of the algorithm involved a weighted adjacency matrix $\mathbf{M}$. This was the same as the FCS matrix described earlier, with a few modifications. The columns of $\mathbf{M}$ were all normalized so that its row elements had a sum of 1. Thus, each column essentially represented the probability distribution of a signal passing through a residue to its neighbours. Furthermore, before the ResidueRank algorithm proceeded, the residue graph was partitioned into its individual connected components. Each component was a connected subgraph within the entire residue network such that no edge existed between any two distinct components. Each connected component was represented by a new weighted adjacency matrix. This separated the residue network into disjoint components such that a signal originating in one component could not transfer to a residue in a different component. For a modified weighted adjacency matrix of a connected component, $\mathbf{M'}$, each non-zero entry $\mathbf{M'_{ij}}$ indicates a directed edge from residue R_i to residue R_j, with both residues being members of the component. $\mathbf{M'}$ is essentially a copy of $\mathbf{M}$, except with all entries involving residues not in the connected component being set to 0. Since each connected component is treated separately, the ResidueRank algorithm takes as input the modified weighted adjacency matrix $\mathbf{M'}$ that represents an individual component. The algorithm iteratively adjusts the ResidueRank weight of each residue according to its outgoing links (correlations with neighbours) and the weights of all residues that link to it for a set number of iterations. In each iteration, we adjusted all the residues' weights at once by premultiplying $\mathbf{M'}$ with the weight vector $\mathbf{v}$. The details of the algorithm, which is adapted from the original PageRank algorithm [74], are shown in Algorithm 2.

Algorithm 2 ResidueRank algorithm

```
    **Input** $N \times N$ modified weighted adjacency matrix $\mathbf{M'}$, component size $K$, number of iterations $iters$
    **Output** weight vector $\mathbf{v}$
1: procedure ResidueRank
2:     $i \leftarrow 1$
3:     $\mathbf{v} \leftarrow$ Vector(size: $N$)
4:     For $i \in [1, N], \mathbf{v_i} \leftarrow \frac{1}{K}$ if $i$ corresponds to a residue in the component
5:     while $i \leq iters$ do
6:         $\mathbf{v} \leftarrow \mathbf{M'v}$                                        ▷ Update $\mathbf{v}$
7:         $i \leftarrow i + 1$
8:     $vSum \leftarrow \sum_{i=1}^{N} \mathbf{v_i}$
9:     For $i \in [1, N], \mathbf{v_i} \leftarrow \frac{K\mathbf{v_i}}{vSum}$        ▷ Scale elements of $\mathbf{v}$ by size of connected component
```

The output weight vector **v** only had non-zero entries in positions corresponding to residues that were part of the connected component. Once **v** was calculated for all the connected components, the weight vectors were summed to produce the final vector holding the ResidueRank weights of every residue in the residue network. Since ResidueRank was run for every connected component of the residue interaction graph, line 8 in the algorithm scales the values in **v** based on the size of the connected component. This ensured that the final weights for residues in smaller connected components were not overrepresented.

3.6.2 The Fluctuation Propagation Algorithm

While the ResidueRank algorithm was able to produce a ranking of the most important residues based on side-chain interactions, it lacked the ability to derive allosteric pathways and did not take into consideration the outside influence of water and other molecules surrounding the protein. To address these shortcomings, we designed a propagation-based algorithm that was an extension of that used by Wang et al. [97]. This algorithm simulated the propagation of a signal starting at some residue, such that propagation from one residue to the next was probabilistically determined by the corresponding FCS value. The higher the FC value, the greater the chance of a signal successfully propagating to the next residue. As the signal could “branch” out and follow multiple pathways simultaneously, it could be described as a moving wavefront that coursed from one start position throughout the rest of the network. When a “branch” of the signal stopped, the residues through which it passed formed a directed path $(R_a, R_b, R_c, ...)$ specifying the order that the residues were visited.

This termination occurred when the signal failed to propagate to an unvisited residue. After all signal paths terminated, there were multiple directed pathways, each representing an individual signal path that began at the signal initiating residue. The process was repeated t times, and the number of times a signal passed through each residue R_i was recorded as p_i. Thus, we were able to calculate the relative frequency of signal propagations through each residue as $\frac{p_i}{t}$. We will refer to this value as the Fluctuation Concentration (FC) of R_i. As with ResidueRank, all the residues within the protein were ranked by their FC values, with a higher FC indicating that a residue is more active in conveying fluctuations. Furthermore, the residues in each signal path were recorded, so important allosteric pathways could be obtained.

To factor in the "dampening" effect of the surrounding water bath, the probability of propagation from a residue R_i to a neighbouring residue R_j was also affected by the solvent accessible surface area of R_i. The greater the amount of surface that was exposed to the water bath, the lower the chance of a signal propagating from R_i to R_j. Thus, a signal propagating towards a residue near the surface of the protein had a lower chance of succeeding compared to a signal propagating towards a residue buried within the protein core.

Propagation itself was implemented using a breadth-first search approach. A propagation would never travel in the reverse direction, since the breadth-first search method meant that a residue that was already visited would not be visited again in the same propagation. The base propagation algorithm with a single signal initiating residue is shown in Algorithm 3.

Algorithm 3 Fluctuation propagation algorithm

```
Input $N \times N$ FCS matrix $\mathbf{A}$, start residue $R_{start}$, number of iterations $iters$
Output FC vector $\mathbf{v}$
procedure Propagate
    $i \leftarrow 1$
    $\mathbf{v} \leftarrow$ Vector(size: $N$)
    while $i \leq iters$ do
        visited $\leftarrow$ Vector(size: $N$)
        Initialize empty queue $\mathbf{q}$
        $\mathbf{q}$.push($r_{start}$)
        while $\mathbf{q}$ is not empty do                          ▷ Perform propagation via BFS
            $r_{curr} \leftarrow \mathbf{q}$.pop()
            for unvisited neighbour residue $R_{nbr}$ of $R_{curr}$ do
                $areaSAS \leftarrow$ solvent accessible surface area of $R_{nbr}$
                $totalSurface \leftarrow$ total surface area of $R_{nbr}$
                $SAScheck \leftarrow 1 - min(1, areaSAS/totalSurface)$
                $p_1 \leftarrow$ random number between [0, 1]
                $p_2 \leftarrow$ random number between [0, 1]
                if $p_1 < SAScheck$ and $p_2 < \mathbf{A_{curr,nbr}}$ then     ▷ Propogate to neighbour
                    $\mathbf{visited_{nbr}} \leftarrow 1$
                    $\mathbf{q}$.push($r_{nbr}$)
        $\mathbf{v} \leftarrow \mathbf{v} + \mathbf{visited}$
        $i \leftarrow i + 1$
    For $i \in [1, N], \mathbf{v_i} \leftarrow \frac{\mathbf{v_i}}{max(v)}$                    ▷ Normalize FC values
```

While this algorithm worked with a single known start residue, there may be cases where the signal initiating residue is not known or there are multiple residues that can each be a signal initiating residue. Thus, a more generalized algorithm was desired.

Rather than propagate from only a single starting residue, we instead performed multiple propagations, each starting from a different residue. The final FC values were calcuated as a weighted average of the FC values from each separate propagation. Performing Algorithm 3 with k different starting residues resulted in a set of FC values $\{FC_1^a, FC_2^a, ..., FC_k^a\}$ for a residue r_a. We then compiled a set of the corresponding starting residues' solvent accessible surface areas $\{SAS_1, SAS_2, ..., SAS_k\}$. The final weighted fluctuation concentration of r_a, FC_a, was calculated as $FC_a = (\sum_{i=1}^{k} SAS_i FC_i^a)/k$. Propagations only started from non-buried residues, as fully buried residues have a solvent accessible surface area of 0. This method aimed to simulate the effects of signal propagations originating from outside molecules, such as water molecules, hitting a residue near the protein's surface. The weighting of FC values by the solvent accessible surface area of the starting residue recognized that residues that are more exposed to the surrounding water bath were more likely to be hit by some molecule and initiate a signal propagation. Algorithm 4 implements this method.

Algorithm 4 Propagation algorithm from all non-buried residues

```
    Input $N \times N$ FCS matrix $\mathbf{A}$, number of iterations $iters$
    Output FC vector $\mathbf{v}^*$
procedure PropagateAll
    $b \leftarrow$ number of non-buried residues
    $\mathbf{v} \leftarrow$ Vector(size: $b$)
    $maxSAS \leftarrow$ max surface accessible surface area among all residues
    for each non-buried residue $R_i$ do
        $residueSAS \leftarrow$ surface acessible surface area of $R_i$
        $\mathbf{v_i} \leftarrow$ Propagate($\mathbf{A}$, $R_i$, $iters$)
        $\mathbf{v_i} \leftarrow (\frac{residueSAS}{maxSAS}) * \mathbf{v_i}$        ▷ Scale FC vector by SAS of start residue
    $\mathbf{v}^* \leftarrow$ mean($\mathbf{v}$)
```

The modified propagation algorithm took into consideration the difference in solvent accessibility across the protein, and so the weighting of propagations in this manner was more likely to reflect the biophysical nature of how allosteric signal propagations are initiated *in vivo*.

3.7 Identifying High Fluctuation Sites

The final step in our computational pipeline was the analysis of propagation results. This involved the organization of the FC data along with comparisons with the residue propagation network. Here, we define the residue propagation network as the edge-weighted and node-weighted digraph whose nodes represent the residues, and the weight of a node represent the FC value of the corresponding residue. The weight of an edge, $w(i, j)$, represents the total frequency that residue R_j was visited from residue R_i during all propagations. In other words, $w(i, j)$ represents the number of times edge (i, j) was traversed, and the weighted in-degree of a node R_i, $in\text{-}deg(i) = \sum_{k=1} w(k, i)$, represents the number of times R_i was visited.

Sorting the FC data gave us the relative fluctuation conveying ability of each residue in the network. Furthermore, we sorted the edges by their weights and the nodes by their in-degree, which gave us the relative frequency of edge traversals among all pathways and the relative activity of each residue, respectively. This information revealed more details about the physical nature of the propagations.

We compared the FC data, edge weights, and node in-degree values for both the apo and holo forms of a protein. This reveals the effects of ligand binding on the signal transmission, such as changes in the fluctuation conveying ability of residues along with the frequency of pathways taken during signal propagation. Examining the differences between the residue propagation networks for these two forms was useful for assessing how secondary structures such as alpha helices and beta sheets affected allosteric mechanisms. If these structures were allosterically important, we expected their presence or absence to affect the propagation results. Thus, we could remove or alter these structures from their native PDB conformations and run the modified protein through the entire computational pipeline again, comparing how the final FC values and residue propagation networks differed from their original states. These insights provide a better understanding of the functional importance of such a structure. Since it is commonly understood that allostery serves as an internal mechanism that facilitates a protein's functionality (interactions with a ligand or another protein), we wanted to uncover the relationship between structural elements and protein functionality.

Chapter 4

Results

In this Chapter, we present results obtained by applying the computational methods described in Chapter 3. We first show the residue networks and FCS values resulting from kernelized canonical correlation analysis on side-chain ensemble samples. Using the FCS data, we then present and compare experimental results derived by applying the ResidueRank and propagation algorithms. The ResidueRank algorithm provided the relative importance for each residue in a residue network, while the propagation algorithms derived a relative importance for each residue and each edge within a residue network.

Our analyses were focused on the PDZ3 domain of the PSD-95 protein (apo form PDB ID: 1BFE, holo form PDB ID: 1BE9) for three reasons: its relatively small size, the abundance of prior experimental and computational research on the protein, and the fact that its backbone atom positions have only very small deviations when comparing its apo and holo forms. The small size of the structures meant that performing the required energy minimizations would take less time than performing the minimizations on a larger structure and visualization of the residue networks would not be as cluttered. Furthermore, since the PDZ domain of PSD-95 is well-studied in literature, we had prior studies to compare our results. The functionality of the PDZ3 domain is also well-documented, as we know that the protein is a scaffold protein. These types of proteins act as regulators in cellular signalling pathways. Some other protein structures that met these two criteria had greater backbone deviations between their apo and holo forms, but our objective was to study allostery from a side-chain centric point of view that assumes a fixed backbone. These characteristics make the PDZ3 domain an especially compatible case study for our methodology, since as mentioned in Chapter 1, our goal was to show that our model is able to reveal allosteric sites in “simpler” protein structures before analyzing more complicated cases.

The PDZ3 domain has a total of 115 residues (numbered 301-415 on chain A), and the ligand in 1BE9 is composed of five residues (numbered 5 to 9 on chain B). This ligand is a C-terminal peptide derived from the CRIPT protein. The active site of the PDZ3 domain is typically considered to be just under the binding site. However, there is some debate regarding how the active site of PDZ3 should be defined, especially when considering the presence of adjacent domains [50].

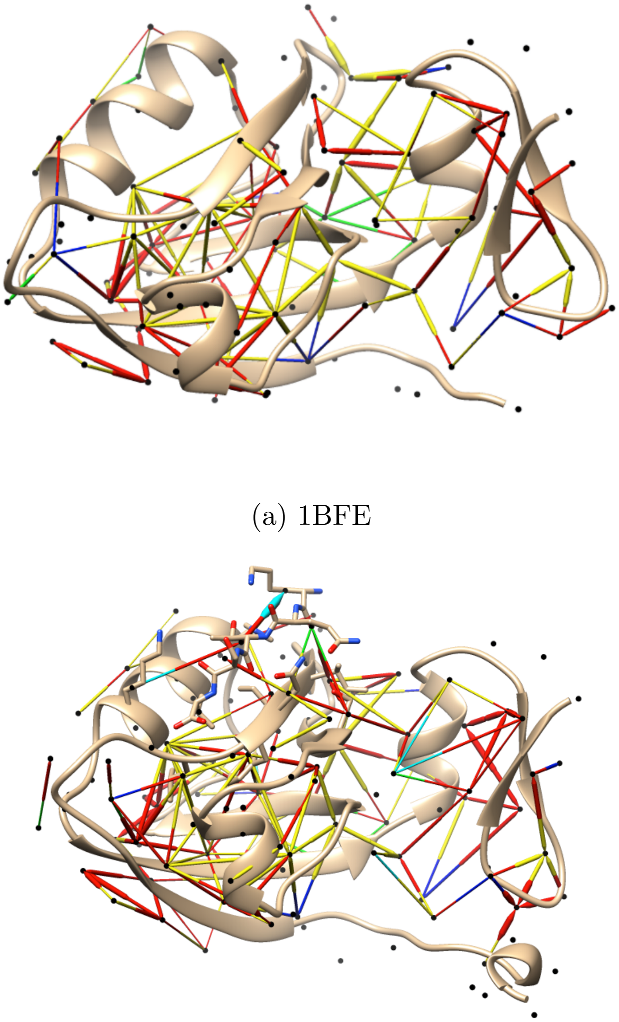
4.1 Fluctuation Coupling Analysis

We generated side-chain conformations using the repeated perturbation and energy minimization method, and then we applied correlation analysis to derive FCS values for each residue and its neighbours. This gave us a residue network where relations between residues are described by their FCS values. Figure 4.1 shows a 3D visual display of the protein structure illustrating the FCS values along with the degrees of freedom of residue-residue interactions between all neighbouring residues for the apo and holo forms of PSD-95.

4.2 ResidueRank Weights

The results of the ResidueRank algorithm on 1BFE and 1BE9 are shown in Table 4.1. The naming convention we use for residues is as follows: <chain ID>_<sequence number>_<type>. For example, A_318_ARG refers to the residue Arg318 on chain A. To validate our results, we compared the rankings with the experimental result from McLaughlin et al. [62]. They performed mutagenesis experiments on PSD-95 to obtain a list of residues with high functional cost of mutation relative to wild-type PSD-95 [62]. The functional cost of a mutation was measured as a loss-of-function or a gain-of-function. The residues with the highest loss-of-function were determined to be the residues with the highest functional cost. For reference, Figure B.1 shows the results of the mutational analyses that McLaughlin et al. performed on PDZ3.

For 1BFE, the highest ranked residues were Ile388, Phe325, Leu353, Ile359, and Leu379. For 1BE9, the highest ranked residues were Leu379, Leu353, Phe325, Ile388, and Ile316. From the top 20 ranked residues for 1BFE, eight were also determined to be functionally important [62]. For 1BE9, this number increased to 11. Overall, the residues with a high ResidueRank weight agree with the experimentally determined functionally important residues. Note that out of the 20 functionally important residues identified by McLaughlin et al., six of these are either alanine, glycine, or proline [62], and as such would not



(a) 1BFE

(b) 1BE9

Figure 4.1: Visual display of (a) 1BFE and (b) 1BE9 with connections showing the FCS and DoF interactions between neighbouring residues. A residue is represented as a black sphere at the centroid of that residue. Interactions between residues are shown with a pair of spindles connecting two residues. For a pair of spindles connecting residues R_i and R_j, the spindle closer to R_i represents the FCS from R_i to R_j, and the spindle closer to R_j represents the FCS from R_j to R_i. The size of the spindle indicates the relative FCS, with a larger spindle indicating a greater FCS. The colour of the spindle indicates the DoF of the closest atom to the neighbouring residue; a DoF of one is shown with a red spindle, a DoF of two is shown with a yellow spindle, a DoF of three is shown with a green spindle, a DoF of four is shown with a cyan spindle, and a DoF of five is shown with a blue spindle.

Table 4.1: Residues with the highest weights after applying Algorithm 2 to 1BFE and 1BE9. Residues in bold font are those that were experimentally identified by McLaughlin et al. to be functionally important [62].

Rank	**1BFE**		**1BE9**	
	Residue	**Weight**	**Residue**	**Weight**
1	**A_388_ILE**	0.0411	**A_379_LEU**	0.0351
2	**A_325_PHE**	0.0369	**A_353_LEU**	0.0342
3	**A_353_LEU**	0.0368	**A_325_PHE**	0.0310
4	**A_359_ILE**	0.0342	**A_388_ILE**	0.0304
5	**A_379_LEU**	0.0325	A_316_ILE	0.0266
6	**A_362_VAL**	0.0305	**A_359_ILE**	0.0257
7	A_386_VAL	0.0302	A_386_VAL	0.0256
8	A_314_ILE	0.0278	**A_323_LEU**	0.0234
9	**A_323_LEU**	0.0273	A_314_ILE	0.0231
10	A_316_ILE	0.0271	**A_362_VAL**	0.0223
11	**A_327_ILE**	0.0268	A_387_THR	0.0191
12	**A_338_ILE**	0.0254	**A_338_ILE**	0.0179
13	A_312_ARG	0.0192	A_315_VAL	0.0178
14	A_350_SER	0.0188	A_337_PHE	0.0176
15	A_404_SER	0.0183	**A_372_HIS**	0.0167
16	A_357_ASP	0.0177	A_326_ASN	0.0162
17	A_401_GLU	0.0176	A_350_SER	0.0160
18	A_397_TYR	0.0175	A_409_SER	0.0159
19	A_318_ARG	0.0170	**A_367_LEU**	0.0157
20	A_315_VAL	0.0168	**A_327_ILE**	0.0156

be included in our analysis. However, ResidueRank is only capable of giving a relative importance ranking of residues. This algorithm was unable to provide information about possible allosteric pathways and did not take into consideration the effects of water or other surrounding molecules. To remedy these issues, we used the propagation-based algorithm.

4.3 Fluctuation Propagation Analysis

We ran the propagation algorithm to obtain the FC of residues. For 1BFE, the residues with the highest FC were Val362, Arg312, Asp357, Val386, and Ile338. For 1BE9, these were Gln358, Thr387, Leu360, Phe337, and Val362. Table 4.2 contains a more substantial list of residues with high FC for these two proteins structures.

One noticeable difference between the FC results for 1BFE and 1BE9 is that for 1BE9, many residues located on or near the α_3 helix had higher FC values than the corresponding residues in 1BFE. The α_3 helix of PDZ3 includes residues 394-399. Out of the 20 residues with the highest FCs for 1BE9, eight were either part of the α_3 helix or were a neighbour of a residue on the α_3 helix, compared to only two residues out of the 20 highest ranked residues for 1BFE. The importance of the α_3 helix in PSD-95 was emphasized in other studies as well [89, 102], so our results agree with the assessment that the α_3 helix plays a role in allostery for PDZ3. For 1BFE, 10 of the top 20 residues ranked by FC were included in the list of functionally important residues identified by McLaughlin et al. [62]. Therefore, the propagation algorithm was more accurate than the ResidueRank algorithm for the identification of allosterically important residues when considering the structure of 1BFE.

During the propagation experiments, we also kept track of the number of times each node and edge was visited. Table 4.3 shows the most frequently visited nodes and Table 4.4 shows the most frequently visited edges for PSD-95. These tables also include results for modified structures of 1BFE and 1BE9 that simulate an inactivation of the α_3 helix. The modification was either a truncation of the helix or a phosphorylation of the Tyr397 residue. We chose these specific changes because, as noted in Table 4.2, the α_3 helix is considered to be allosterically important when a ligand is bound to PSD-95 [52]. Furthermore, the phosphorylation of Tyr397 was shown to modulate binding affinity for PSD-95 [102].

Organizing the node and edge traversal frequencies revealed which physical locations in the protein structure experienced the greatest amount of signal propagations. The more frequently traversed edges indicate pathways that a signal is likelier to take. We also compared the node visit frequencies with the FC rankings. The FC rankings and node

Table 4.2: Residues with the highest fluctuation concentration (FC) after applying Algorithm 4 to the apo form (1BFE) and the holo form (1BE9) of the PDZ3 domain. Residues in bold font are those that were experimentally identified by McLaughlin et al. to be functionally important. Residues in italics font are either part of the α_3 helix or a neighbour of an α_3 helix residue.

Rank	**1BFE**		**1BE9**	
	Residue	**FC**	**Residue**	**FC**
1	**A_362_VAL**	1.000	A_358_GLN	1.00
2	A_312_ARG	0.992	A_387_THR	0.982
3	A_357_ASP	0.990	A_360_LEU	0.947
4	A_386_VAL	0.985	*A_337_PHE*	0.945
5	**A_338_ILE**	0.964	**A_362_VAL**	0.932
6	**A_388_ILE**	0.963	**A_328_ILE**	0.908
7	A_318_ARG	0.948	*A_396_GLU*	0.907
8	**A_325_PHE**	0.946	*A_393_LYS*	0.902
9	**A_327_ILE**	0.944	*A_397_TYR*	0.901
10	**A_353_LEU**	0.939	*A_400_PHE*	0.898
11	*A_392_TYR*	0.934	A_339_SER	0.888
12	**A_323_LEU**	0.925	*A_412_ILE*	0.871
13	**A_379_LEU**	0.909	A_368_ARG	0.862
14	A_350_SER	0.899	A_357_ASP	0.855
15	A_316_ILE	0.898	A_363_ASN	0.852
16	*A_412_ILE*	0.884	A_312_ARG	0.850
17	A_314_ILE	0.881	A_386_VAL	0.849
18	A_363_ASN	0.870	*A_392_TYR*	0.848
19	**A_367_LEU**	0.869	*A_394_PRO*	0.847
20	**A_359_ILE**	0.865	A_334_GLU	0.845

Table 4.3: Most frequently visited nodes from applying the propagation algorithm to several protein structures. A ”-T” suffix indicates a modified structure that is truncated at the α_3 helix (including residues 394-415). A ”-P” suffix indicates a modified structure in which Tyr397 is phosphorylated.

Node Rank	**1BFE**	**1BFE-T**	**1BFE-P**	**1BE9**	**1BE9-T**	**1BE9-P**
1	A_357_ASP	A_357_ASP	A_357_ASP	A_362_VAL	A_362_VAL	A_362_VAL
2	A_312_ARG	A_362_VAL	A_312_ARG	A_392_TYR	A_386_VAL	A_386_VAL
3	A_362_VAL	A_386_VAL	A_362_VAL	A_357_ASP	A_327_ILE	A_357_ASP
4	A_386_VAL	A_312_ARG	A_386_VAL	A_327_ILE	A_387_THR	A_392_TYR
5	A_338_ILE	A_316_ILE	A_338_ILE	A_386_VAL	A_388_ILE	A_327_ILE
6	A_325_PHE	A_388_ILE	A_325_PHE	A_412_ILE	A_323_LEU	A_388_ILE
7	A_316_ILE	A_325_PHE	A_388_ILE	A_388_ILE	A_379_LEU	A_338_ILE
8	A_353_LEU	A_338_ILE	A_316_ILE	A_338_ILE	A_338_ILE	A_323_LEU
9	A_388_ILE	A_327_ILE	A_323_LEU	A_353_LEU	A_353_LEU	A_353_LEU
10	A_327_ILE	A_323_LEU	A_353_LEU	A_312_ARG	A_325_PHE	A_387_THR
11	A_392_TYR	A_353_LEU	A_327_ILE	A_323_LEU	A_357_ASP	A_312_ARG
12	A_323_LEU	A_379_LEU	A_392_TYR	A_387_THR	A_359_ILE	A_325_PHE
13	A_350_SER	A_350_SER	A_314_ILE	A_325_PHE	A_367_LEU	A_379_LEU
14	A_314_ILE	A_314_ILE	A_350_SER	A_379_LEU	A_318_ARG	A_363_ASN
15	A_412_ILE	A_318_ARG	A_379_LEU	A_359_ILE	A_363_ASN	A_359_ILE
16	A_379_LEU	A_359_ILE	A_359_ILE	A_318_ARG	A_312_ARG	A_367_LEU
17	A_359_ILE	A_392_TYR	A_412_ILE	A_316_ILE	A_316_ILE	A_318_ARG
18	A_318_ARG	A_367_LEU	A_318_ARG	A_363_ASN	A_314_ILE	A_316_ILE
19	A_354_ARG	A_363_ASN	A_367_LEU	A_404_SER	A_365_VAL	A_365_VAL
20	A_307_ILE	A_354_ARG	A_354_ARG	A_367_LEU	A_336_ILE	A_314_ILE

Table 4.4: Most frequently visited edges from applying the propagation algorithm to several proteins. A "-T" suffix indicates a modified structure that is truncated at the alpha helix 3 (including residues 394-415), and a "-P" suffix indicates a modified structure in which Tyr397 is phosphorylated.

Edge Rank	1BFE	1BFE-T	1BFE-P	1BE9	1BE9-T	1BE9-P
1	A_392_TYR → A_412_ILE	A_357_ASP → A_392_TYR	A_392_TYR → A_412_ILE	A_338_ILE → A_341_ILE	A_357_ASP → A_392_TYR	A_338_ILE → A_341_ILE
2	A_338_ILE → A_341_ILE	A_338_ILE → A_341_ILE	A_338_ILE → A_341_ILE	A_392_TYR → A_307_ILE	A_338_ILE → A_341_ILE	A_389_ILE → A_361_SER
3	A_357_ASP → A_392_TYR	A_327_ILE → A_372_HIS	A_357_ASP → A_392_TYR	A_392_TYR → A_354_ARG	A_389_ILE → A_361_SER	A_387_THR → A_315_VAL
4	A_392_TYR → A_307_ILE	A_392_TYR → A_307_ILE	A_392_TYR → A_307_ILE	A_398_SER → A_406_VAL	A_363_ASN → A_387_THR	A_357_ASP → A_392_TYR
5	A_327_ILE → A_372_HIS	A_314_ILE → A_352_GLU	A_327_ILE → A_372_HIS	A_404_SER → A_398_SER	A_387_THR → A_315_VAL	A_392_TYR → A_307_ILE
6	A_314_ILE → A_352_GLU	A_323_LEU → A_318_ARG	A_314_ILE → A_352_GLU	A_389_ILE → A_361_SER	A_362_VAL → A_365_VAL	A_392_TYR → A_354_ARG
7	A_323_LEU → A_318_ARG	A_363_ASN → A_387_THR	A_353_LEU → A_350_SER	A_363_ASN → A_387_THR	A_392_TYR → A_307_ILE	A_392_TYR → A_412_ILE
8	A_353_LEU → A_350_SER	A_353_LEU → A_357_ASP	A_327_ILE → A_336_ILE	A_314_ILE → A_352_GLU	A_314_ILE → A_352_GLU	A_363_ASN → A_387_THR
9	A_412_ILE → A_355_LYS	A_386_VAL → A_363_ASN	A_398_SER → A_406_VAL	A_387_THR → A_315_VAL	A_353_LEU→ A_357_ASP	A_362_VAL → A_365_VAL
10	A_327_ILE → A_336_ILE	A_318_ARG → A_384_GLN	A_323_LEU → A_318_ARG	A_357_ASP → A_392_TYR	A_392_TYR → A_354_ARG	A_337_PHE → A_334_GLU
11	A_388_ILE → A_386_VAL	A_327_ILE → A_336_ILE	A_359_ILE → A_367_LEU	A_392_TYR → A_412_ILE	A_387_THR → A_385_THR	A_314_ILE → A_352_GLU
12	A_359_ILE → A_367_LEU	A_353_LEU → A_350_SER	A_388_ILE → A_386_VAL	A_362_VAL → A_365_VAL	A_318_ARG → A_321_THR	A_387_THR → A_385_THR
13	A_363_ASN → A_387_THR	A_359_ILE → A_367_LEU	A_386_VAL → A_363_ASN	A_409_SER → A_306_ASP	A_379_LEU → A_336_ILE	A_409_SER → A_306_ASP
14	A_412_ILE → A_404_SER	A_312_ARG → A_354_ARG	A_412_ILE → A_404_SER	A_325_PHE → A_327_ILE	A_387_THR → A_389_ILE	A_398_SER → A_406_VAL
15	A_386_VAL → A_363_ASN	A_388_ILE → A_386_VAL	A_318_ARG → A_384_GLN	A_353_LEU → A_350_SER	A_386_VAL → A_318_ARG	A_353_LEU → A_350_SER
16	A_318_ARG → A_384_GLN	A_362_VAL → A_367_LEU	A_363_ASN → A_387_THR	A_412_ILE → A_404_SER	A_327_ILE → A_372_HIS	A_337_PHE → A_400_PHE
17	A_404_SER → A_398_SER	A_353_LEU → A_312_ARG	A_412_ILE → A_398_SER	A_337_PHE → A_358_GLN	A_325_PHE → A_327_ILE	A_404_SER → A_398_SER
18	A_388_ILE → A_362_VAL	A_387_THR → A_315_VAL	A_353_LEU → A_357_ASP	A_387_THR → A_385_THR	A_372_HIS → B_7_THR	A_325_PHE → A_327_ILE
19	A_392_TYR → A_354_ARG	A_314_ILE → A_312_ARG	A_404_SER → A_401_GLU	A_357_ASP → A_312_ARG	A_314_ILE → A_312_ARG	A_386_VAL → A_318_ARG
20	A_353_LEU → A_357_ASP	A_388_ILE → A_362_VAL	A_312_ARG → A_354_ARG	A_318_ARG → A_321_THR	A_358_GLN → A_337_PHE	A_353_LEU → A_357_ASP

visit frequencies for 1BFE showed high similarities. However, when comparing the FC rankings and node visit frequencies for 1BE9, there were more significant differences. For example, Gln358 had the highest FC value for 1BE9, but it did not appear in the 20 most visited nodes. Neither did Leu360 or Phe337, which were the residues with the 3rd and 4th highest FC values, respectively. These discrepancies indicate that a residue with a high fluctuation activity does not necessarily have an equally high FC value.

Figures 4.2 and 4.3 show 2D network representations of the propagation networks for 1BFE and 1BE9, respectively. The clustering of the residues reveals those residues that experienced a heavier concentration of fluctuations. For 1BFE, the main cluster of residues contains those that are near the binding site, and their importance is reflected by their higher FC values and greater concentration of edge connections compared to residues with a lower degree. For 1BE9, the clustering is less apparent, though there appears to be two main clusters: one cluster contains residues near the binding site, and one cluster contains residues at or near the α_3 helix.

The edges connecting the two main clusters in the 1BE9 network did not show higher than average usage, and thus the increased fluctuation concentration exhibited by residues at the α_3 helix was not due to propagations from the ligand binding site to the α_3 helix. Our model suggests that the functional importance of the α_3 helix is not due to a single signal pathway from the binding site to the helix. Instead, signals that are initiated from highly solvent accessible residues will travel to the α_3 helix.

We examined the effect that the solvent accessible surface area of the propagation-initiating residues had on the fluctuation concentration results. Figures 4.4-4.9 shows the contribution that the solvent accessibility of each starting residue of Algorithm 4 has on the final FC values for PSD-95.

Many residues had a consistent FC regardless of the location of the signal initiating residue. The residues with the highest weighted average FC values had consistently high FCs over all the propagations, and residues with low weighted average FC values had consistently low FCs. The consistency of FC values indicates that, in general, the results of a propagation from any residue are likely to be representative of the final weighted FC values. There were some outliers in this regard, but those outliers did not have a significant effect on the final FC values. For example, for 1BFE, Glu373 and Gln374 had large solvent accessible surface areas. Propagations initiated from those two residues also tended to terminate early, as only residues close by, such as Ser371 and Ile377, had significant FC values. However, the final averaged FC values for those residues were still fairly low, so the outlier propagations starting from Glu373 and Gln374 did not significantly impact the final FC values.

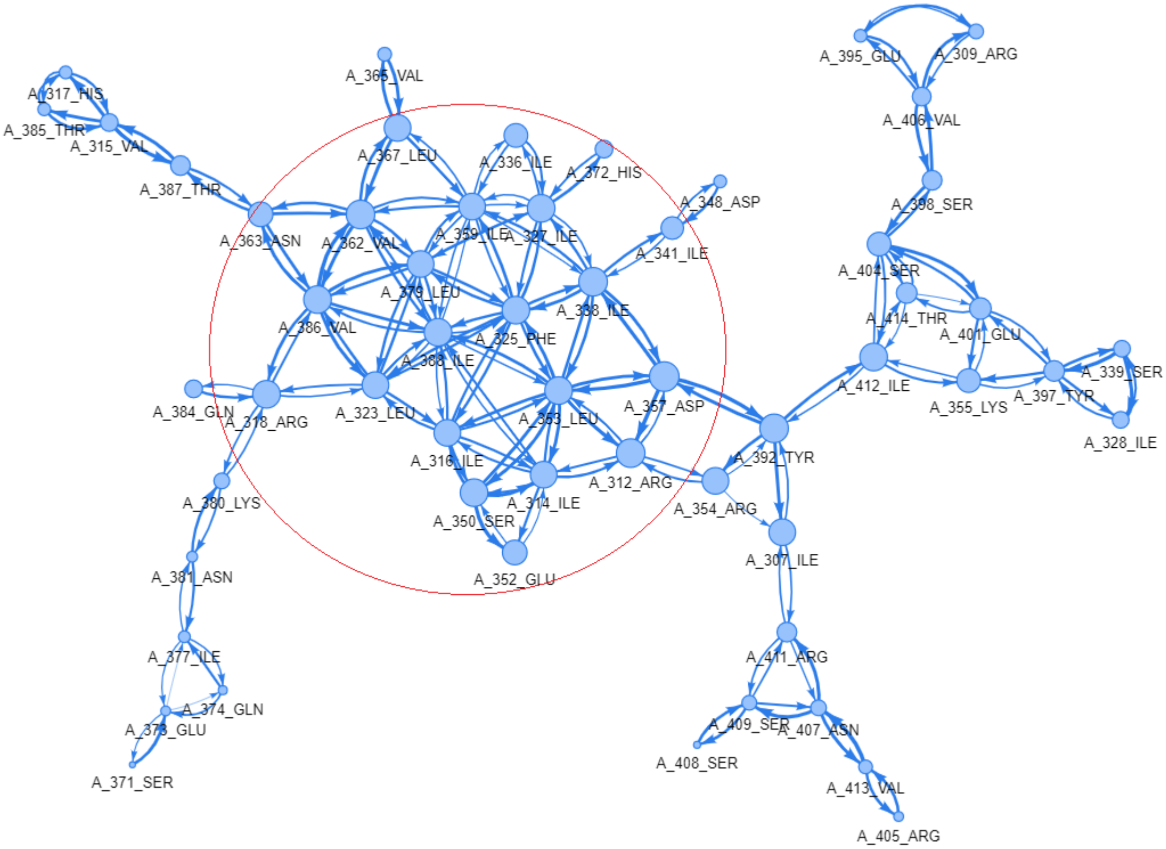


Figure 4.2: 2D visual displays of the residue propagation networks for 1BFE. The size of each node is proportional to the FC value of its corresponding residue. The width of each edge is proportional to how frequently that edge was traversed during propagations. The area of the graph circled in red indicates the concentration of residues near the binding site. This graph was created using the Pyvis library for Python [78].

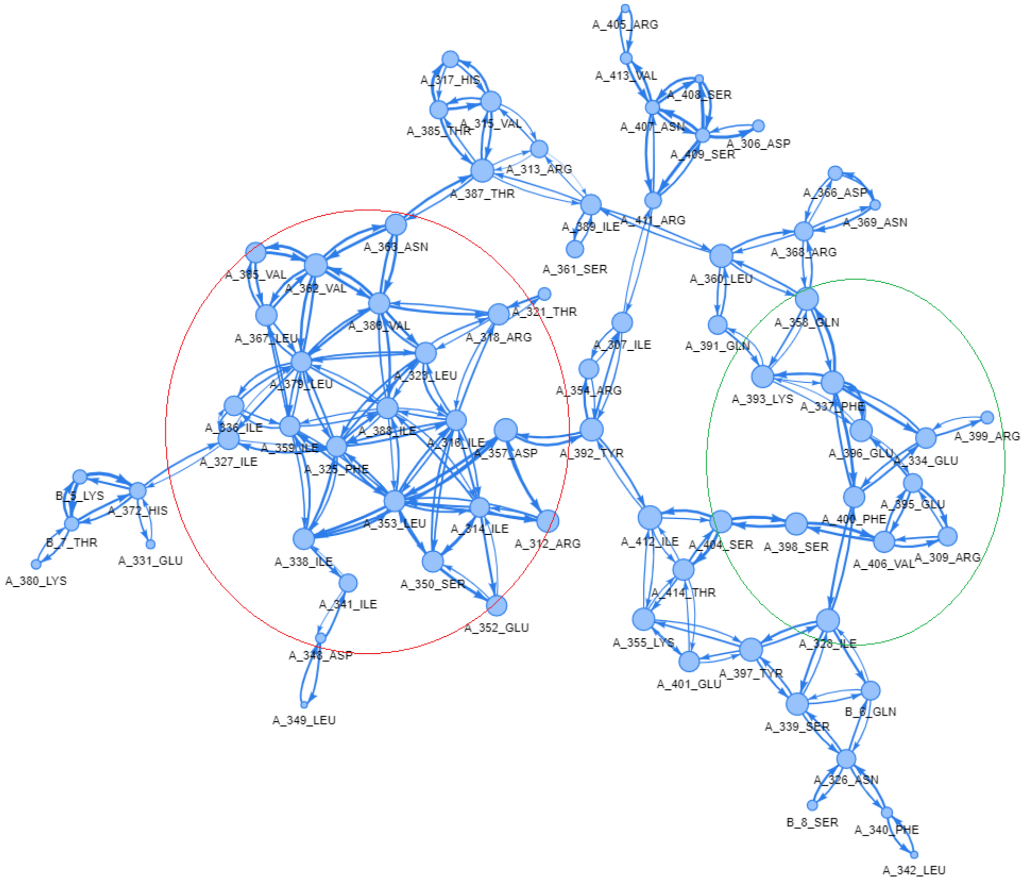


Figure 4.3: 2D visual displays of the residue propagation networks for 1BE9. The size of each node is proportional to the FC value of its corresponding residue. The width of each edge is proportional to how frequently that edge was traversed during propagations. The area of the graph circled in red indicates the concentration of residues near the binding site, and the area circled in green indicates the concentration of residues near the α_3 helix. This graph was created using the Pyvis library for Python [78].

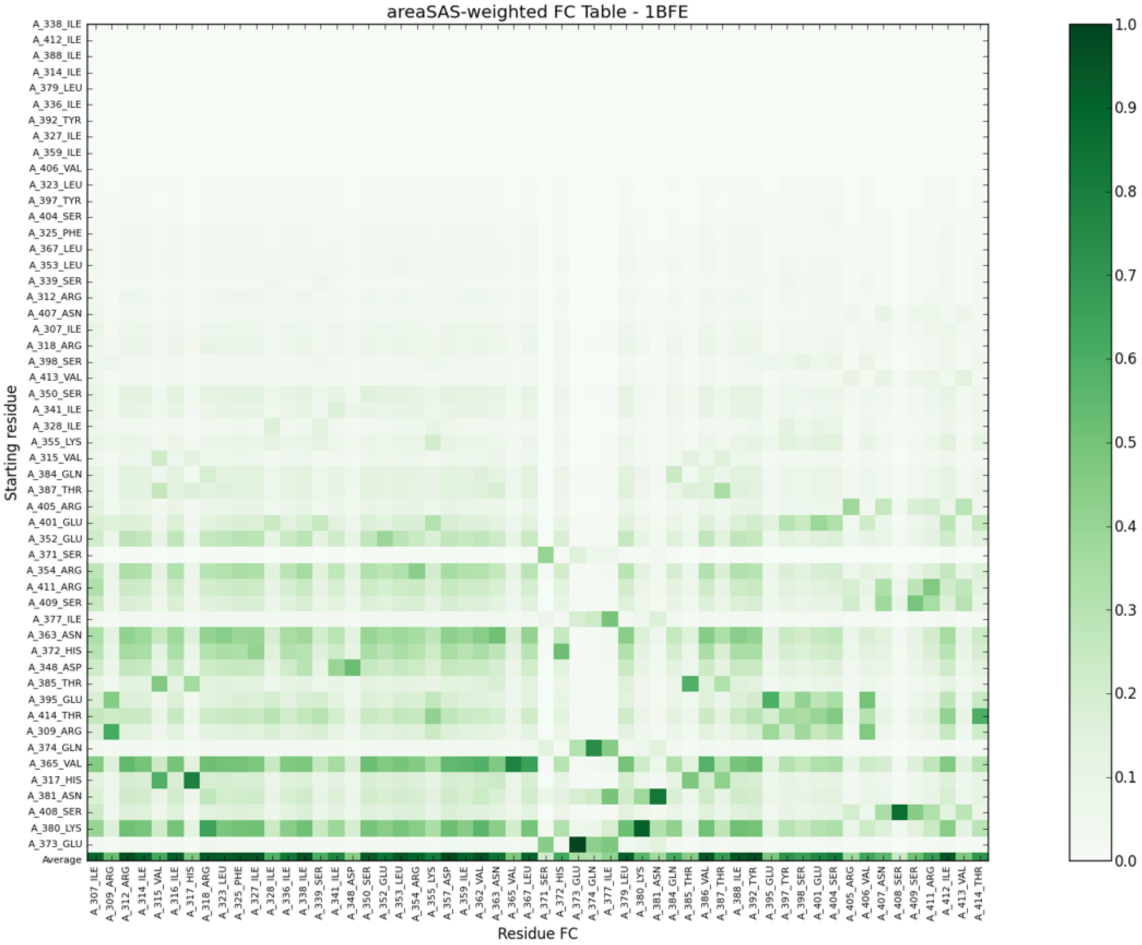


Figure 4.4: Heat map showing the FC propagations for 1BFE, weighted by the solvent accessible surface area of each signal initiating residue. The variable areaSAS represents the solvent accessible surface area.

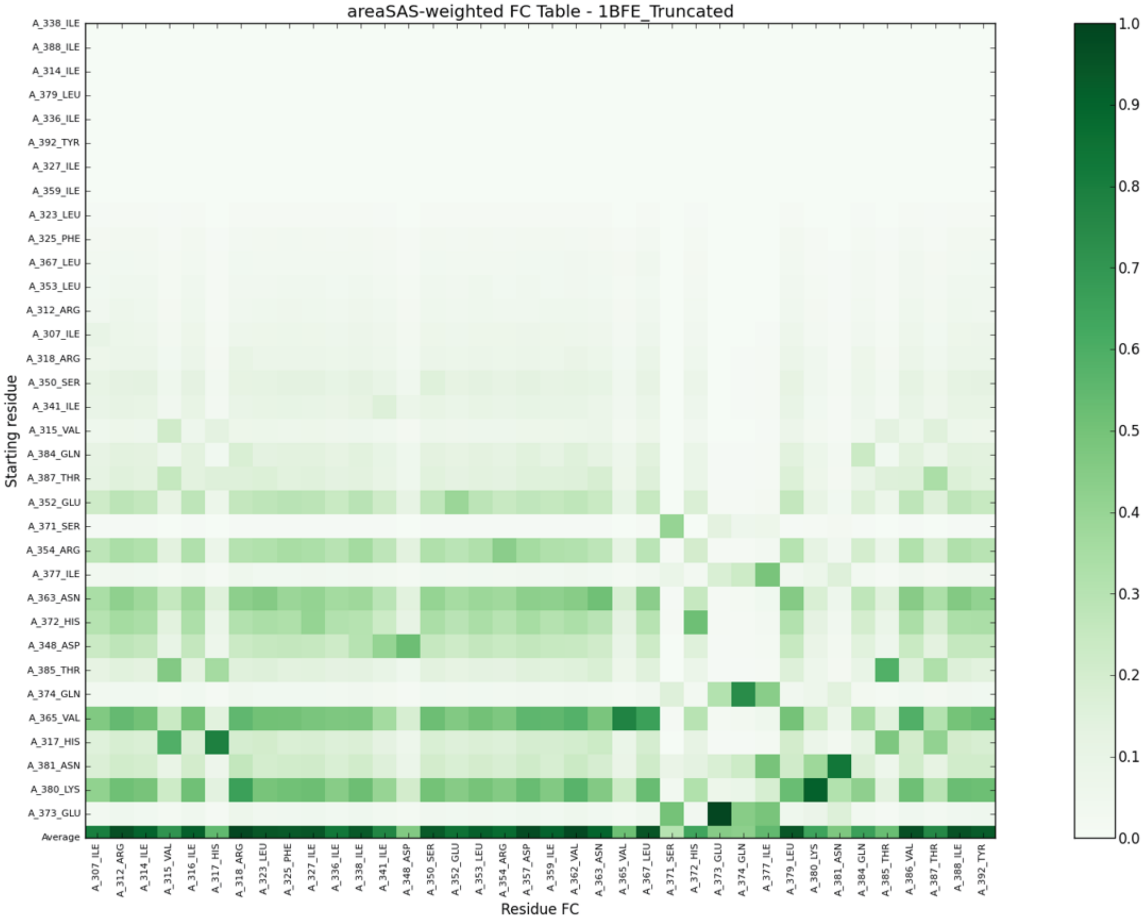


Figure 4.5: Heat map showing the FC propagations for 1BFE truncated at the α_3 helix (residues 394-415), weighted by the solvent accessible surface area of each signal initiating residue. The variable areaSAS represents the solvent accessible surface area.

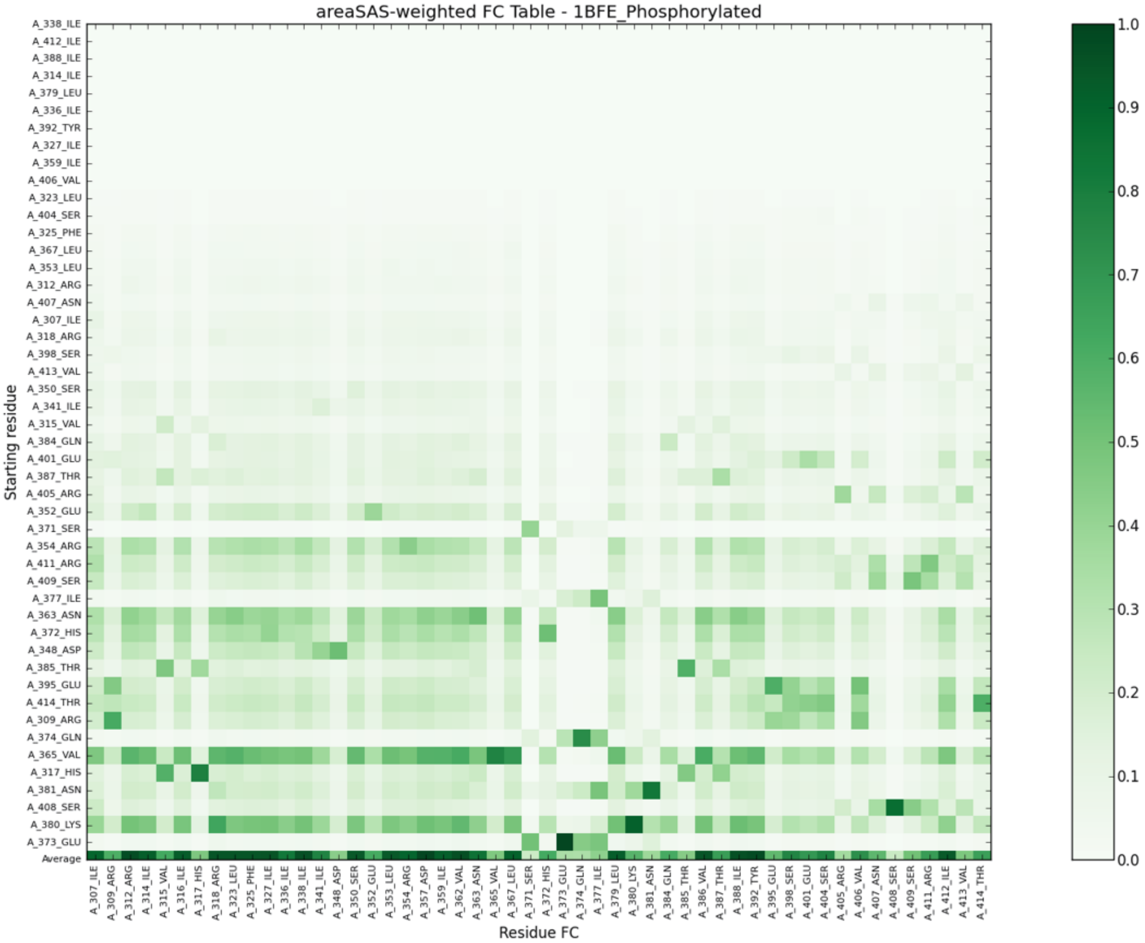


Figure 4.6: Heat map showing the FC propagations for 1BFE phosphorylated at Tyr397, weighted by the solvent accessible surface area of each signal initiating residue. The variable areaSAS represents the solvent accessible surface area.

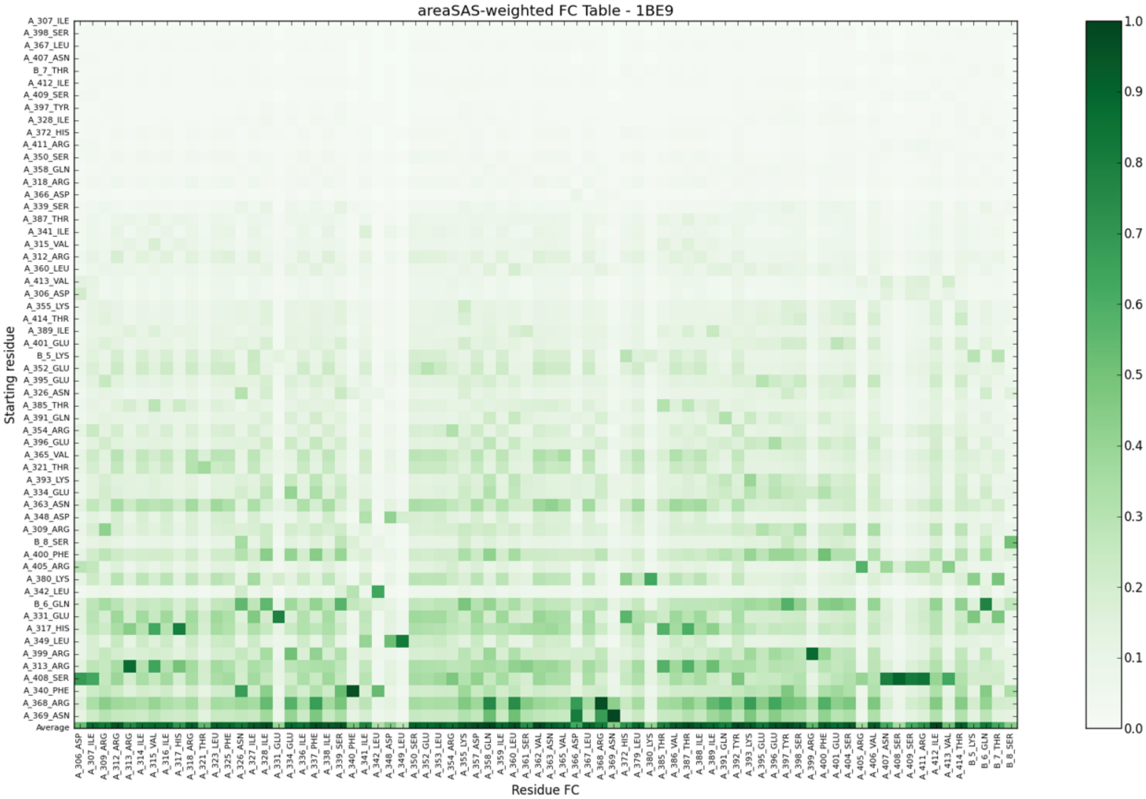


Figure 4.7: Heat map showing the FC propagations for 1BE9, weighted by the solvent accessible surface area of each signal initiating residue. The variable areaSAS represents the solvent accessible surface area.

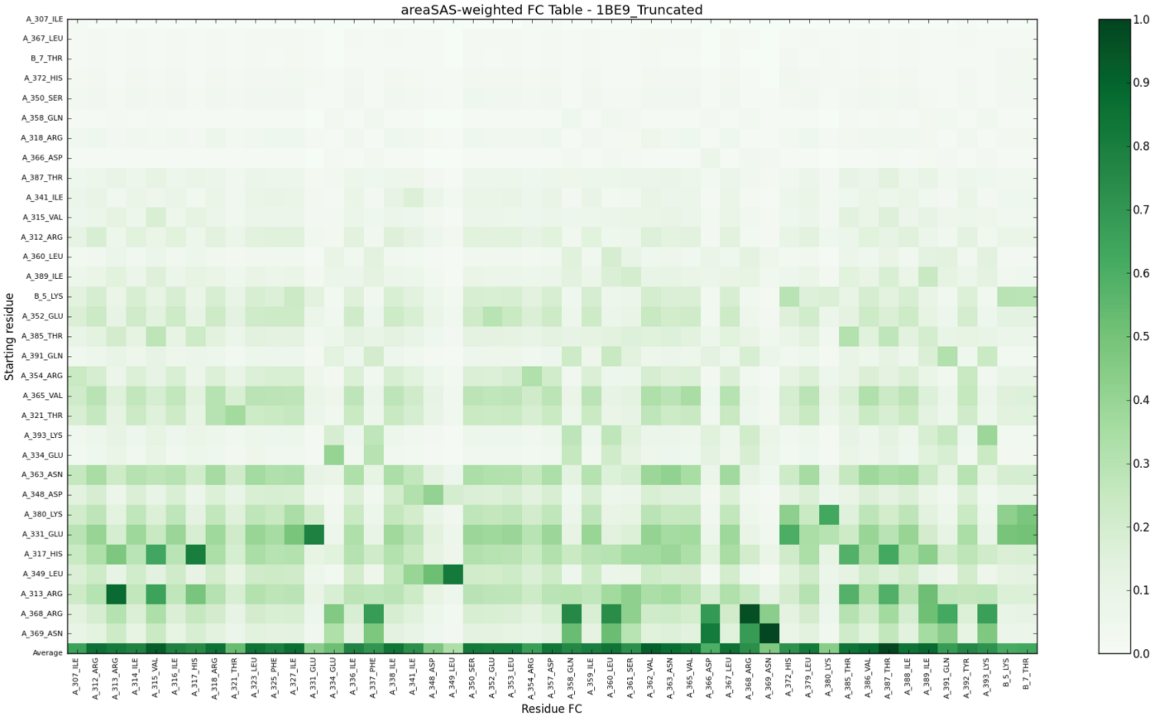


Figure 4.8: Heat map showing the FC propagations for 1BFE truncated at the α_3 helix (residues 394-415), weighted by the solvent accessible surface area of each signal initiating residue. The variable areaSAS represents the solvent accessible surface area.

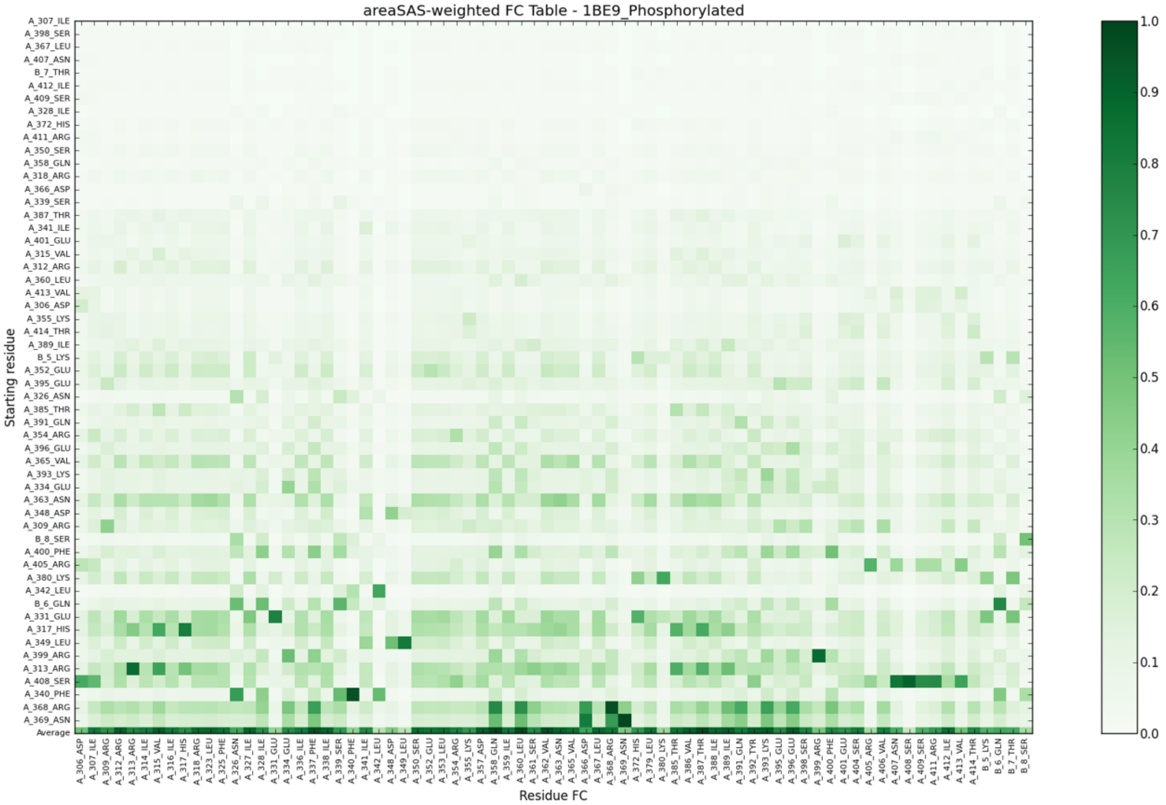


Figure 4.9: Heat map showing the FC propagations for 1BFE phosphorylated at Tyr397, weighted by the solvent accessible surface area of each signal initiating residue. The variable areaSAS represents the solvent accessible surface area.

We also note that phosphorylation of Tyr397 did not significantly change the FC results. For both 1BFE and 1BE9, the FC values from the base structures and the phosphorylated structures were very similar. When we ran the propagations on the structures truncated at the α_3 helix, there were more noticeable differences in the FC values across the propagations. These differences were more apparent when comparing the results for 1BE9 (shown in Figure 4.7) and the truncated structure of 1BE9 (shown in Figure 4.8). In the weighted averaged FC row for 1BE9, the residues with the highest FCs have very similar values. However, in the weighted averaged FC row for 1BE9-T, the residue with the highest FC is easily seen to be Thr387, while the rest of the residues have distinctly lower FCs.

Chapter 5

Discussion and Conclusions

In Figure 4.1, we show the degrees of freedom of atom-atom interactions based on the closest pair of side-chain atoms between neighbouring residues in PDZ3. About 80% of the interactions were between atoms with one degrees of freedom (DoF1) or two degrees of freedom (DoF2). This observation is in line with the sampling of DoF interactions shown in Table 3.1, as DoF1:DoF1 and DoF1:DoF2 interactions were by far the most common types of interactions between neighbouring residues. From Table 3.1, we see that DoF1:DoF1, DoF1:DoF2, and DoF2:DoF2 interactions form about 84% of all DoF interactions. The degrees of freedom of interactions are important because the degrees of freedom of an atom dictates its range of motion, and our model is based on side-chain interactions. From a signal propagation point of view, we theorized that a signal transfer involving a low DoF atom interacting with a higher DoF atom would result in a dissipation of the signal's strength. The loss of signal strength would be due to an atom with a lower range of motion colliding with an atom with a higher range of motion. Thus, we expected that allosteric signals would be primarily conveyed through DoF1:DoF1 interactions.

Most of the spindles outside of the core of PDZ3 are red in Figure 4.1, indicating that DoF1 interactions were the most common in those areas. In particular, interactions around the α_3 helix were almost entirely composed of DoF1:DoF1 interactions for both 1BFE and 1BE9. There were more DoF2:DoF2 interactions by the ligand binding site, but we note that the majority of interactions still involved DoF1 atoms. Furthermore, in DoF1:DoF2 interactions, the FCS of the DoF1 → DoF2 direction was generally higher than that of the reverse direction. There were not many interactions involving DoF3, DoF4, or DoF5 atoms in PDZ3. The prevalence of DoF1:DoF1 interactions in PDZ3, especially around the surface of the protein where a signal is likely to be initiated from contacts with a water molecule or some other outside molecule, supports the notion that interactions involving

DoF1 atoms are crucial for a model of allostery based on side-chain interactions. The imbalance of FCS values in DoF1:DoF2 interactions also suggests that a dissipation of signal strength occurs in DoF1:DoF2 interactions.

Our decision to implement a propagation algorithm in addition to the ResidueRank algorithm was primarily because the ResidueRank algorithm lacked a relevant biophysical basis. While the prospect of applying a well-known mathematical algorithm to allostery was exciting, the results by themselves were not convincing enough to conclude that our ResidueRank algorithm provided a definitive model for allosteric behaviour. The original PageRank algorithm was devised for an entirely different purpose, and it lacked the nuances needed to explain allostery in a protein system. We also found that when running the ResidueRank algorithm with equal weights for all interacting residues (meaning that all non-zero entries in $\mathbf{M}'$ had the same value), the ranking of residues was similar to what we obtained with the "regular" weights (results for equal weights not shown).

Unlike ResidueRank, the propagation algorithms shown in Algorithms 3 and 4 presented a relevant biophysical explanation for allostery. Furthermore, this type of propagation has been explored in previous computational studies on allostery, such as in the Ohm model [97]. Our model had several modifications, such as the assumption that a signal can be initiated at any solvent exposed residue rather than just a single residue, the use of fluctuation couplings to determine propagation probabilities, and the influence of solvent exposed surface area on propagation probabilities. These differences meant that the propagations in our model represented a more realistic version of protein signal propagations compared to the Ohm model, which did not account for the influence of the surrounding water bath in protein systems.

Based on the prevalence of residues close to the α_3 helix in the list of residues with high FCs for 1BE9 (Table 4.2), we wanted to explore the structural differences between 1BFE and 1BE9. The main goal of the phosphorylation and truncation modifications to PDZ3 was to see if the propagation results between the two modified structures would be similar. Similarities between the results would indicate that a phosphorylation of Tyr397 was functionally similar to an inactivation of the α_3 helix as a whole. We also wanted to compare the propagations between the baseline structures for PDZ3 (the unmodified structures of 1BFE and 1BE9) and the structures with an inactivated α_3 helix (the modified structures of 1BFE and 1BE9).

Table 4.3, Table 4.4, and Figures 4.4-4.9 suggest that a truncation of the α_3 helix had a greater impact on the propagations than just a phosphorylation of Tyr397. Experimental results have confirmed that Tyr397 phosphorylation regulates the ligand binding affinity of the PDZ3 domain of PSD-95 [102], so we expected that a truncation of the α_3 helix and a phosphorylation of Tyr397 would have similar effects. In Table 4.3, we observe that the ranking of the most frequently visited nodes differ substantially between the baseline structures (1BFE/1BE9) and the truncated structures (1BFE-T/1BE9-T). A similar observation is made with the ranking of the most frequently visited edges in Table 4.4. However, the ranking of the nodes and edges of the phosphorylated structures (1BFE-P/1BE9-P) had more similarities with the ranking of the nodes and edges of the baseline structures of PDZ3. For example, in Table 4.4, the most frequently used edge for propagations with 1BFE and 1BFE-P is Tyr392 → Ile412, and the most frequently used edge for propagations with 1BE9 and 1BE9-P is Ile338 → Ile341. In contrast, the most frequently used edge for both 1BFE-T and 1BE9-T is Asp357 → Tyr392. Looking at the solvent accessible surface area weighted FC values in Figures 4.4-4.9 also shows more visible differences between the truncated structures and the baseline structures.

We were unable to find an existing PDB file that contained a structure of the PDZ3 domain of PSD-95 with Tyr397 already phosphorylated, so we instead manually derived a phosphorylated structure using Chimera. We created the phosphorylated structure by replacing the hydroxyl group in Tyr397 with a phosphate group and then performing an energy minimization to stabilize the protein structure. However, it is possible that this method did not create an accurate *in vivo* representation of PDZ3 with a phosphorylated Tyr397 residue. Another explanation is that since the energy minimizations were not the same as MD simulations, they were unable to characterize the functional effects of a single residue phosphorylation. After all, the minimizations only moved the molecular system towards a local energy minimum, and would not take into account as many factors as a full-scale MD simulation.

5.1 Limitations of This Study

Our model assumed that allosteric signals are conveyed through side-chain fluctuations. While this assumption was reasonable for proteins that have minimal differences in backbone positions between their apo and holo forms (such as the PDZ3 domain of PSD-95), our model may not work as well with other proteins that have more drastic differences in backbone atom positions. For example, calmodulin is a small protein that is frequently used for studying allostery, but due to the extensive conformational changes between its apo and holo forms, our model would not be able to accurately characterize the allosteric processes that result in these changes. For these proteins that have more noticeable backbone fluctuations, it can be argued that the backbone plays a greater role alongside side-chains in allosteric communications.

Another limitation was the computational cost of the energy minimization procedures in Chimera. These minimizations were akin to a restricted molecular dynamics simulation, so while the computational cost was not as high as a full-scale MD simulation, the minimizations were still the slowest step in our computational pipeline. For reference, performing all the minimizations for the unmodified structure of 1BE9 took around 50 hours on a laptop with an Intel i7-11800H@2.30GHz processor and 16GB of RAM. However, the timeframe to perform the necessary energy minimizations was still in the range of several days for most proteins, compared to the multiple weeks it would take an MD simulation to cover an allosteric process from start to end. Furthermore, the main purpose of the energy minimizations was to generate a sample of plausible protein conformations, so the amount of time spent on performing energy minimizations for a single protein structure could be closely estimated by the number of conformations needed and the time spent on a single energy minimization on that protein structure.

5.2 Future Work

The most obvious extension would be to apply our model to other known allosteric proteins with minimal backbone perturbations. Since we have validated our results with experimentally determined residues that have a high functional cost for the PDZ3 domain of PSD-95, the next step would be to see if our model agrees with experimental results obtained for similar proteins. Another extension would be to apply our model to allosteric proteins with more significant backbone perturbations. Here, the goal would be to verify whether a side-chain centric model can accurately characterize allosteric behaviour in a protein system with both backbone and side-chain fluctuations. A computational model of allostery that can be quickly applied to any protein would be an invaluable tool in protein research.

Our model could also be improved by adding the ability to derive allosteric pathways. While the main goal of our model was to identify specific residues that were allosterically important, the ability to derive directed pathways of residues would also be useful. Each pathway would represent the path a signal would likely take when propagating from one site in the protein structure to a different site. The main reason we did not prioritize allosteric pathway generation was that we considered the hypothesis that a single putative pathway would not accurately characterize the behaviour of an allosteric signal in many cases. Even within our model, a propagation consists of not just a single pathway, but rather a set of multiple directed pathways that each have a common starting node. Especially when we consider a series of propagations over multiple starting nodes (residues), a single pathway may not accurately reflect the nature of the propagations. However, we could use the edge frequencies to derive a list of frequently travelled “incomplete” pathways. Each pathway here would not represent the full path from a signal initiating site to an allosteric site, but instead part of a path that an allosteric signal would likely take during a propagation.

5.3 Conclusions

In summary, we developed a side-chain centric model of allostery and applied it to the PDZ3 domain of the PSD-95 protein. Through the use of a network-based propagation algorithm that simulated the transmission of a signal in a protein structure, we identified key residues and secondary structure elements, such as side-chain fluctuations within a helix, for allosteric signalling. We also examined the behaviour of propagations within our model. The advantage of a network model is that it depends only on the protein structure, bypassing the high computational costs of molecular dynamics simulations. Our model also does not disregard the contribution of side-chains often omitted in NMA and ENMs. By analyzing side-chain fluctuations to simulate the propagation of a signal, our method was rooted in a biophysically relevant model of allostery. Overall, our results agreed with experimental studies that identified functionally important residues and secondary structure elements within the PDZ3 domain of PSD-95.

References

[1] Alakent, B., and Ince, Z. N. G. *Elucidating Allosteric Communication in Proteins via Computational Methods*, vol. 3. Bentham Science Publishers: Sharjah, UAE, 2017.

[2] Alexandrov, V., Lehnert, U., Echols, N., Milburn, D., Engelman, D., and Gerstein, M. Normal modes for predicting protein motions: a comprehensive database assessment and associated web tool. *Protein Science 14*, 3 (2005), 633–643.

[3] Bach, F. R., and Jordan, M. I. Kernel independent component analysis. *Journal of Machine Learning Research 3*, Jul (2002), 1–48.

[4] Bahar, I., Atilgan, A. R., and Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design 2*, 3 (1997), 173–181.

[5] Bahar, I., Lezon, T. R., Bakan, A., and Shrivastava, I. H. Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chemical Reviews 110*, 3 (2010), 1463–1497.

[6] Bahar, I., and Rader, A. Coarse-grained normal mode analysis in structural biology. *Current Opinion in Structural Biology 15*, 5 (2005), 586–592.

[7] Bai, F., Branch, R. W., Nicolau Jr, D. V., Pilizota, T., Steel, B. C., Maini, P. K., and Berry, R. M. Conformational spread as a mechanism for cooperativity in the bacterial flagellar switch. *Science 327*, 5966 (2010), 685–689.

[8] Bauer, J. A., Pavlović, J., and Bauerová-Hlinková, V. Normal mode analysis as a routine part of a structural investigation. *Molecules 24*, 18 (2019), 3293.

[9] Beauchamp, K. A., Bowman, G. R., Lane, T. J., Maibaum, L., Haque, I. S., and Pande, V. S. MSMBuilder2: modeling conformational dynamics on the picosecond to millisecond scale. *Journal of Chemical Theory and Computation 7*, 10 (2011), 3412–3419.

[10] Bertaccini, E. J., Trudell, J. R., and Lindahl, E. Normal-mode analysis of the glycine alpha1 receptor by three separate methods. *Journal of chemical information and modeling 47*, 4 (2007), 1572–1579.

[11] Bounova, G., and De Weck, O. Overview of metrics and their correlation patterns for multiple-metric topology analysis on heterogeneous graph ensembles. *Physical Review E 85*, 1 (2012), 016117.

[12] Bowman, G. R., Pande, V. S., and Noé, F. *An introduction to Markov state models and their application to long timescale molecular simulation*, vol. 797. Springer Science & Business Media, 2013.

[13] Bozovic, O., Zanobini, C., Gulzar, A., Jankovic, B., Buhrke, D., Post, M., Wolf, S., Stock, G., and Hamm, P. Real-time observation of ligand-induced allosteric transitions in a PDZ domain. *Proceedings of the National Academy of Sciences 117*, 42 (2020), 26031–26039.

[14] Burkowski, F. J. *Computational and Visualization Techniques for Structural Bioinformatics Using Chimera*. CRC Press, 2014.

[15] Censoni, L., dos Santos Muniz, H., and Martínez, L. A network model predicts the intensity of residue-protein thermal coupling. *Bioinformatics 33*, 14 (2017), 2106–2113.

[16] Clenshaw, C. W. Chebyshev series for mathematical functions. *NPL Mathematical Tables 5* (1962).

[17] Cohen, M., Potapov, V., and Schreiber, G. Four distances between pairs of amino acids provide a precise description of their interaction. *PLoS Computational Biology 5*, 8 (2009), e1000470.

[18] Collier, G., and Ortiz, V. Emerging computational approaches for the study of protein allostery. *Archives of Biochemistry and Biophysics 538*, 1 (2013), 6–15.

[19] Cooper, A., and Dryden, D. Allostery without conformational change. *European Biophysics Journal 11*, 2 (1984), 103–109.

[20] Daily, M. D., and Gray, J. J. Local motions in a benchmark of allosteric proteins. *Proteins: Structure, Function, and Bioinformatics 67*, 2 (2007), 385–399.

[21] Daily, M. D., and Gray, J. J. Allosteric communication occurs via networks of tertiary and quaternary motions in proteins. *PLoS Computational Biology 5*, 2 (2009), e1000293.

[22] Daily, M. D., Upadhyaya, T. J., and Gray, J. J. Contact rearrangements form coupled networks from local motions in allosteric proteins. *Proteins: Structure, Function, and Bioinformatics 71*, 1 (2008), 455–466.

[23] Doerr, S., Harvey, M., Noé, F., and De Fabritiis, G. HTMD: high-throughput molecular dynamics for molecular discovery. *Journal of Chemical Theory and Computation 12*, 4 (2016), 1845–1852.

[24] Dror, R. O., Dirks, R. M., Grossman, J., Xu, H., and Shaw, D. E. Biomolecular simulation: a computational microscope for molecular biology. *Annu. Rev. Biophys 41*, 1 (2012), 429–452.

[25] Duke, T., Le Novere, N., and Bray, D. Conformational spread in a ring of proteins: a stochastic approach to allostery. *Journal of Molecular Biology 308*, 3 (2001), 541–553.

[26] Dyson, H. J., Wright, P. E., and Scheraga, H. A. The role of hydrophobic interactions in initiation and propagation of protein folding. *Proceedings of the National Academy of Sciences 103*, 35 (2006), 13057–13061.

[27] Edward, W. Y., and Koshland, D. E. Propagating conformational changes over long (and short) distances in proteins. *Proceedings of the National Academy of Sciences 98*, 17 (2001), 9517–9520.

[28] Emekli, U., Schneidman-Duhovny, D., Wolfson, H. J., Nussinov, R., and Haliloglu, T. HingeProt: automated prediction of hinges in protein structures. *Proteins: Structure, Function, and Bioinformatics 70*, 4 (2008), 1219–1227.

[29] Flores, S., Echols, N., Milburn, D., Hespenheide, B., Keating, K., Lu, J., Wells, S., Yu, E. Z., Thorpe, M., and Gerstein, M. The database of macromolecular motions: new features added at the decade mark. *Nucleic Acids Research 34*, suppl_1 (2006), D296–D301.

[30] Gadiyaram, V., Dighe, A., Ghosh, S., and Vishveshwara, S. Network rewiring during allostery and protein-protein interactions: A graph spectral approach. *Allostery* (2021), 89–112.

[31] Gerek, Z. N., and Ozkan, S. B. Change in allosteric network affects binding affinities of PDZ domains: analysis through perturbation response scanning. *PLoS Computational Biology 7*, 10 (2011), e1002154.

[32] Goncearenco, A., Mitternacht, S., Yong, T., Eisenhaber, B., Eisenhaber, F., and Berezovsky, I. N. SPACER: server for predicting allosteric communication and effects of regulation. *Nucleic Acids Research 41*, W1 (2013), W266–W272.

[33] Greener, J. G., and Sternberg, M. J. AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. *BMC Bioinformatics 16*, 1 (2015), 1–7.

[34] Greener, J. G., and Sternberg, M. J. Structure-based prediction of protein allostery. *Current Opinion in Structural Biology 50* (2018), 1–8.

[35] Gunasekaran, K., Ma, B., and Nussinov, R. Is allostery an intrinsic property of all dynamic proteins? *Proteins: Structure, Function, and Bioinformatics 57*, 3 (2004), 433–443.

[36] Harbrecht, H., Peters, M., and Schneider, R. On the low-rank approximation by the pivoted Cholesky decomposition. *Applied Numerical Mathematics 62*, 4 (2012), 428–440.

[37] Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation 16*, 12 (2004), 2639–2664.

[38] Hertig, S., Latorraca, N. R., and Dror, R. O. Revealing atomic-level mechanisms of protein allostery with molecular dynamics simulations. *PLoS Computational Biology 12*, 6 (2016), e1004746.

[39] Hilser, V. J. An ensemble view of allostery. *Science 327*, 5966 (2010), 653–654.

[40] Hilser, V. J., and Thompson, E. B. Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proceedings of the National Academy of Sciences 104*, 20 (2007), 8311–8315.

[41] Hilser, V. J., Wrabl, J. O., and Motlagh, H. N. Structural and energetic basis of allostery. *Annual Review of Biophysics 41* (2012), 585–609.

[42] Hinsen, K. Analysis of domain motions by approximate normal mode calculations. *Proteins: Structure, Function, and Bioinformatics 33*, 3 (1998), 417–429.

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

[44] Isralewitz, B., Gao, M., and Schulten, K. Steered molecular dynamics and mechanical functions of proteins. *Current Opinion in Structural Biology 11*, 2 (2001), 224–230.

[45] Issa, N. T., Badiavas, E. V., and Schürer, S. Research techniques made simple: Molecular docking in dermatology-a foray into in silico drug discovery. *Journal of Investigative Dermatology 139*, 12 (2019), 2400–2408.

[46] Itoh, K., and Sasai, M. Statistical mechanics of protein allostery: roles of backbone and side-chain structural fluctuations. *The Journal of Chemical pPhysics 134*, 12 (2011), 03B618.

[47] Kannan, N., and Vishveshwara, S. Identification of side-chain clusters in protein structures by a graph spectral method. *Journal of Molecular Biology 292*, 2 (1999), 441–464.

[48] Keegan, M., Siegelmann, H. T., Rietman, E. A., Klement, G. L., and Tuszynski, J. A. Gibbs free energy, a thermodynamic measure of protein–protein interactions, correlates with neurologic disability. *BioMedInformatics 1*, 3 (2021), 201–210.

[49] Laskowski, R. A., Gerick, F., and Thornton, J. M. The structural basis of allosteric regulation in proteins. *FEBS Letters 583*, 11 (2009), 1692–1698.

[50] Laursen, L., Kliche, J., Gianni, S., and Jemth, P. Supertertiary protein structure affects an allosteric network. *Proceedings of the National Academy of Sciences 117*, 39 (2020), 24294–24304.

[51] Leach, K., Sexton, P. M., and Christopoulos, A. Allosteric GPCR modulators: taking advantage of permissive receptor pharmacology. *Trends in Pharmacological Sciences 28*, 8 (2007), 382–389.

[52] Lee, A. L. Contrasting roles of dynamics in protein allostery: NMR and structural studies of CheY and the third PDZ domain from PSD-95. *Biophysical Reviews 7*, 2 (2015), 217–226.

[53] Levitt, M., Sander, C., and Stern, P. S. The normal modes of a protein: Native bovine pancreatic trypsin inhibitor. *International Journal of Quantum Chemistry 24*, S10 (1983), 181–199.

[54] Li, W., Wang, W., and Takada, S. Energy landscape views for interplays among folding, binding, and allostery of calmodulin domains. *Proceedings of the National Academy of Sciences 111*, 29 (2014), 10550–10555.

[55] Liu, J., and Nussinov, R. Allostery: an overview of its history, concepts, methods, and applications. *PLoS Computational Biology 12*, 6 (2016), e1004966.

[56] Lockless, S. W., and Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science 286*, 5438 (1999), 295–299.

[57] Long, D., and Bruschweiler, R. Atomistic kinetic model for population shift and allostery in biomolecules. *Journal of the American Chemical Society 133*, 46 (2011), 18999–19005.

[58] Ma, B., Kumar, S., Tsai, C.-J., and Nussinov, R. Folding funnels and binding mechanisms. *Protein Engineering 12*, 9 (1999), 713–720.

[59] Ma, J. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure 13*, 3 (2005), 373–380.

[60] Materese, C. K., Goldmon, C. C., and Papoian, G. A. Hierarchical organization of eglin c native state dynamics is shaped by competing direct and water-mediated interactions. *Proceedings of the National Academy of Sciences 105*, 31 (2008), 10659–10664.

[61] McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., Schwantes, C. R., Wang, L.-P., Lane, T. J., and Pande, V. S. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophysical Journal 109*, 8 (2015), 1528–1532.

[62] McLaughlin Jr, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S., and Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature 491*, 7422 (2012), 138–142.

[63] Mei, M., and Tan, H. Data expression and protection of intellectual property education resources based on machine learning. *Complexity 2021* (2021).

[64] Meirovitch, H. Recent developments in methodologies for calculating the entropy and free energy of biological systems by computer simulation. *Current Opinion in Structural Biology 17*, 2 (2007), 181–186.

[65] Meirovitch, H., Cheluvaraja, S., and White, R. P. Methods for calculating the entropy and free energy and their application to problems involving protein flexibility and ligand binding. *Current Protein and Peptide Science 10*, 3 (2009), 229–243.

[66] Motlagh, H. N., Wrabl, J. O., Li, J., and Hilser, V. J. The ensemble nature of allostery. *Nature 508*, 7496 (2014), 331–339.

[67] Neyts, E. C., and Bogaerts, A. Combining molecular dynamics with Monte Carlo simulations: implementations and applications. In *Theoretical Chemistry in Belgium*. Springer, 2014, pp. 277–288.

[68] Noguti, T., and Gō, N. Collective variable description of small-amplitude conformational fluctuations in a globular protein. *Nature 296*, 5859 (1982), 776–778.

[69] Nussinov, R., and Tsai, C.-J. Allostery in disease and in drug discovery. *Cell 153*, 2 (2013), 293–305.

[70] Nussinov, R., Tsai, C.-J., and Liu, J. Principles of allosteric interactions in cell signaling. *Journal of the American Chemical Society 136*, 51 (2014), 17692–17701.

[71] Ovchinnikov, V., and Karplus, M. Analysis and elimination of a bias in targeted molecular dynamics simulations of conformational transitions: application to calmodulin. *The Journal of Physical Chemistry B 116*, 29 (2012), 8584–8603.

[72] O'Connor, C. M., Adams, J. U., and Fairman, J. Essentials of cell biology. *Cambridge, MA: NPG Education 1* (2010), 54.

[73] Paci, E., and Karplus, M. Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. *Journal of Molecular Biology 288*, 3 (1999), 441–459.

[74] Page, L., Brin, S., Motwani, R., and Winograd, T. The PageRank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab, 1999.

[75] Panjkovich, A., and Daura, X. Exploiting protein flexibility to predict the location of allosteric sites. *BMC Bioinformatics 13*, 1 (2012), 1–12.

[76] Panjkovich, A., and Daura, X. PARS: a web server for the prediction of protein allosteric and regulatory sites. *Bioinformatics 30*, 9 (2014), 1314–1315.

[77] Paquet, E., and Viktor, H. L. Molecular dynamics, Monte Carlo simulations, and Langevin dynamics: a computational review. *BioMed Research International 2015* (2015).

[78] Perrone, G., Unpingco, J., and Lu, H.-m. Network visualizations with Pyvis and VisJS. *arXiv preprint arXiv:2006.04951* (2020).

[79] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry 25*, 13 (2004), 1605–1612.

[80] Reichheld, S. E., Yu, Z., and Davidson, A. R. The induction of folding cooperativity by ligand binding drives the allosteric response of tetracycline repressor. *Proceedings of the National Academy of Sciences 106*, 52 (2009), 22263–22268.

[81] Rodriguez, G. J., Yao, R., Lichtarge, O., and Wensel, T. G. Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. *Proceedings of the National Academy of Sciences 107*, 17 (2010), 7787–7792.

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

[83] Schrank, T. P., Bolen, D. W., and Hilser, V. J. Rational modulation of conformational fluctuations in adenylate kinase reveals a local unfolding mechanism for allostery and functional adaptation in proteins. *Proceedings of the National Academy of Sciences 106*, 40 (2009), 16984–16989.

[84] Sengupta, U., and Strodel, B. Markov models for the elucidation of allosteric regulation. *Philosophical Transactions of the Royal Society B: Biological Sciences 373*, 1749 (2018), 20170178.

[85] Shapovalov, M. V., and Dunbrack Jr, R. L. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure 19*, 6 (2011), 844–858.

[86] Skjaerven, L., Martinez, A., and Reuter, N. Principal component and normal mode analysis of proteins; a quantitative comparison using the groel subunit. *Proteins: Structure, Function, and Bioinformatics 79*, 1 (2011), 232–243.

[87] Śledź, P., and Caflisch, A. Protein structure-based drug design: from docking to molecular dynamics. *Current Opinion in Structural Biology 48* (2018), 93–102.

[88] Soltan Ghoraie, L., Burkowski, F., and Zhu, M. Using kernelized partial canonical correlation analysis to study directly coupled side chains and allostery in small G proteins. *Bioinformatics 31*, 12 (2015), i124–i132.

[89] Stevens, A. O., and He, Y. Allosterism in the PDZ family. *International Journal of Molecular Sciences 23*, 3 (2022), 1454.

[90] Taylor, C. C., Mardia, K. V., Di Marzio, M., and Panzera, A. Validating protein structure using kernel density estimates. *Journal of Applied Statistics 39*, 11 (2012), 2379–2388.

[91] Tirion, M. M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical review letters 77*, 9 (1996), 1905.

[92] Tsai, C.-J., Del Sol, A., and Nussinov, R. Allostery: absence of a change in shape does not imply that allostery is not at play. *Journal of Molecular Biology 378*, 1 (2008), 1–11.

[93] Tsai, C.-J., Ma, B., and Nussinov, R. Folding and binding cascades: shifts in energy landscapes. *Proceedings of the National Academy of Sciences 96*, 18 (1999), 9970–9972.

[94] Tsai, C.-J., and Nussinov, R. A unified view of "how allostery works". *PLoS Computational Biology 10*, 2 (2014), e1003394.

[95] Van Den Bedem, H., Bhabha, G., Yang, K., Wright, P. E., and Fraser, J. S. Automated identification of functional dynamic contact networks from x-ray crystallography. *Nature Methods 10*, 9 (2013), 896–902.

[96] Wang, G., and Dunbrack Jr, R. L. PISCES: a protein sequence culling server. *Bioinformatics 19*, 12 (2003), 1589–1591.

[97] Wang, J., Jain, A., McDonald, L. R., Gambogi, C., Lee, A. L., and Dokholyan, N. V. Mapping allosteric communications within individual proteins. *Nature Communications 11*, 1 (2020), 1–13.

[98] Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. Development and testing of a general Amber force field. *Journal of Computational Chemistry 25*, 9 (2004), 1157–1174.

[99] Winger, M., Trzesniak, D., Baron, R., and van Gunsteren, W. F. On using a too large integration time step in molecular dynamics simulations of coarse-grained molecular models. *Physical Chemistry Chemical Physics 11*, 12 (2009), 1934–1941.

[100] Wong, K.-B., and Daggett, V. Barstar has a highly dynamic hydrophobic core: evidence from molecular dynamics simulations and nuclear magnetic resonance relaxation data. *Biochemistry 37*, 32 (1998), 11182–11192.

[101] Yang, L.-W., Rader, A., Liu, X., Jursa, C. J., Chen, S. C., Karimi, H. A., and Bahar, I. o GNM: Online computation of structural dynamics using the Gaussian network model. *Nucleic Acids Research 34*, suppl_2 (2006), W24–W31.

[102] Zhang, J., Petit, C. M., King, D. S., and Lee, A. L. Phosphorylation of a PDZ domain extension modulates binding affinity and interdomain interactions in postsynaptic density-95 (PSD-95) protein, a membrane-associated guanylate kinase (MAGUK). *Journal of Biological Chemistry 286*, 48 (2011), 41776–41785.

[103] Zheng, W., and Doniach, S. A comparative study of motor-protein motions by using a simple elastic-network model. *Proceedings of the National Academy of Sciences 100*, 23 (2003), 13253–13258.

[104] Zhuravlev, P. I., Materese, C. K., and Papoian, G. A. Deconstructing the native state: energy landscapes, function, and dynamics of globular proteins. *The Journal of Physical Chemistry B 113*, 26 (2009), 8800–8812.

[105] Zhuravlev, P. I., and Papoian, G. A. Protein functional landscapes, dynamics, allostery: a tortuous path towards a universal theoretical framework. *Quarterly Reviews of Biophysics 43*, 3 (2010), 295–332.

[106] Zong, C., Papoian, G. A., Ulander, J., and Wolynes, P. G. Role of topology, nonadditivity, and water-mediated interactions in predicting the structures of α/β proteins. *Journal of the American Chemical Society 128*, 15 (2006), 5168–5176.

APPENDICES

Appendix A

Identifying Interacting Residues

A.1 Four-Distances Strategy

When using the four-distances strategy to identify neighbouring residues, as many as two atoms were taken from one residue and as many as two atoms were taken from the other residue to calculate the distances. The sets of atoms used for each possible residue pairing are shown below in Table A.1.

Table A.1: Atoms used for the four-distances neighbour strategy. The first two atoms in each cell correspond to the row residue and the last two atoms correspond to the column residue. Only half the table is filled to prevent redundancies.

	ARG	ASN	ASP	CYS	GLN	GLU	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ARG	CG-NH2-CD-NH2	CG-NH2-OD1-ND2	NH1-NH2-OD1-OD2	CG-NH2-CB-SG	CG-NH2-OE1-NE2	CD-NH2-OE1-OE2	CG-NH2-ND1-NE2	CG-NH2-CG2-CD1	CG-NH2-CD1-CD2	NH1-NH2-CB-NZ	CG-NH2-SD-CE	CG-NH2-CE1-CE2	NH1-NH2-CB-CD	CG-NH2-CB-OG	CG-NH2-CG2-OG1	CG-NH2-CD1-CH2	CG-NH2-CE1-OH	CG-NH2-CG2-CG1
ASN		CB-OD1-OD1-ND2	OD1-ND2-OD1-OD2	CB-OD1-CB-SG	OD1-ND2-OE1-NE2	OD1-ND2-OE1-OE2	OD1-ND2-ND1-NE2	OD1-ND2-CG2-CD1	CB-ND2-CD1-CD2	OD1-ND2-CG-NZ	CB-ND2-CB-CE	OD1-ND2-CG-CZ	OD1-ND2-CB-CD	OD1-ND2-CB-OG	OD1-ND2-CG2-OG1	CB-ND2-NE1-CZ3	OD1-ND2-CE1-OH	CB-ND2-CG2-CG1
ASP			CB-OD2-OD1-OD2	CB-OD1-CB-SG	OD1-OD2-OE1-NE2	OD1-OD2-OE1-OE2	OD1-OD2-ND1-NE2	CB-OD2-CG2-CD1	CB-OD2-CD1-CD2	OD1-OD2-CG-NZ	CB-OD2-CG-CE	CB-OD1-CE1-CE2	CB-OD1-CB-CD	OD1-OD2-CB-OG	OD1-OD2-CG2-OG1	CB-OD1-NE1-CH2	OD1-OD2-CE1-OH	CB-OD2-CG2-CG1
CYS				CB-SG-CB-SG	CB-SG-CB-NE2	CB-SG-CG-OE1	CB-SG-ND1-NE2	CB-SG-CG2-CD1	CB-SG-CD1-CD2	CB-SG-CB-CE	CB-SG-CG-CE	CB-SG-CE1-CE2	CB-SG-CB-CD	CB-SG-CB-OG	CB-SG-CG2-OG1	CB-SG-CB-CH2	CB-SG-CE1-OH	CB-SG-CG2-CG1
GLN					CG-OE1-OE1-NE2	OE1-NE2-OE1-OE2	OE1-NE2-ND1-NE2	CB-NE2-CG2-CD1	CB-NE2-CD1-CD2	OE1-NE2-CG-NZ	CB-NE2-CG-CE	CB-NE2-CE1-CE2	OE1-NE2-CB-CD	OE1-NE2-CB-OG	OE1-NE2-CG2-OG1	CB-NE2-NE1-CH2	OE1-NE2-CE1-OH	CB-NE2-CG2-CG1
GLU						CG-OE2-OE1-OE2	OE1-OE2-ND1-NE2	CB-OE2-CG2-CD1	CG-OE1-CD1-CD2	OE1-OE2-CD-NZ	CG-OE1-CG-CE	CB-OE2-CE1-CE2	OE1-OE2-CB-CG	OE1-OE2-CB-OG	OE1-OE2-CG2-OG1	CG-OE1-NE1-CH2	OE1-OE2-CE1-OH	CB-OE2-CG2-CG1
HIS							ND1-NE2-ND1-NE2	ND1-NE2-CG2-CD1	ND1-NE2-CD1-CD2	ND1-NE2-CG-NZ	ND1-NE2-SD-CE	ND1-NE2-CE1-CE2	ND1-NE2-CB-CD	ND1-NE2-CB-OG	ND1-NE2-CG2-OG1	ND1-NE2-CD1-CH2	ND1-NE2-CE1-OH	ND1-NE2-CG2-CG1
ILE								CB-CD1-CG2-CD1	CG2-CD1-CD1-CD2	CG2-CD1-CB-CE	CG2-CD1-SD-CE	CG2-CD1-CE1-CE2	CG2-CD1-CB-CD	CG2-CD1-CB-OG	CG2-CD1-CG2-OG1	CG2-CD1-CB-CH2	CG2-CD1-CE1-OH	CG2-CD1-CG2-CG1
LEU									CG-CD1-CD1-CD2	CD1-CD2-CB-CE	CD1-CD2-SD-CE	CD1-CD2-CE1-CE2	CD1-CD2-CB-CD	CD1-CD2-CB-OG	CD1-CD2-CG2-OG1	CD1-CD2-CB-CH2	CD1-CD2-CE1-OH	CD1-CD2-CG2-CG1
LYS										CB-NZ-CG-NZ	CB-CE-SD-CE	CB-CE-CE1-CE2	CB-NZ-CB-CD	CB-NZ-CB-OG	CG-NZ-CG2-OG1	CB-CE-CD1-CH2	CG-NZ-CE1-OH	CB-CE-CG2-CG1
MET											CG-CE-SD-CE	SD-CE-CE1-CE2	SD-CE-CB-CD	CG-CE-CB-OG	CG-CE-CG2-OG1	CG-CE-CB-CH2	CG-CE-CE1-OH	CG-CE-CG2-CG1
PHE												CE1-CE2-CG-CZ	CE1-CE2-CB-CD	CE1-CE2-CB-OG	CE1-CE2-CG2-OG1	CE1-CE2-CB-CH2	CE1-CE2-CE1-OH	CE1-CE2-CG2-CG1
PRO													CB-CG-CB-CD	CB-CD-CB-OG	CB-CD-CG2-OG1	CB-CD-CD1-CH2	CB-CD-CE1-OH	CB-CD-CG2-CG1
SER														CB-OG-CB-OG	CB-OG-CG2-OG1	CB-OG-CD1-CH2	CB-OG-CE1-OH	CB-OG-CG2-CG1
THR															CB-CG2-CG2-OG1	CG2-OG1-CD1-CH2	CG2-OG1-CE1-OH	CG2-OG1-CG2-CG1
TRP																CB-CH2-CD1-CH2	CD1-CH2-CE1-OH	CB-CH2-CG2-CG1
TYR																	CE1-OH-CE1-OH	CE1-OH-CG2-CG1
VAL																		CB-CG2-CG2-CG1

A.2 A Side-chain Atlas

The side-chain atlas was a tool we developed to help analyze side-chain interactions. The atlas consists of tens of thousands of isolated interacting residue pairs taken from a sample of high-resolution protein structures in the PDB. The list of proteins we sampled can be accessed from the Dunbrack lab's PISCES server [96]. To generate all the necessary pairwise residue conformations, we generated a list of neighbours for each residue in a protein, and stored the coordinates of each residue pair in a PDB file. Neighbours were identified as having C_β atoms within a distance of 5Å. Note that this method of determining residue neighbourhoods is different than that used in Appendix A.1. Residue pairs were separated into categories based on their pair types (e.g. alanine-valine). When storing residue pairs in the atlas, atom coordinates were specified with respect to a coordinate system defined by backbone atoms in the origin residue, so the ordering of the origin-neighbour relationship was important. Using a common coordinate system for residues of the same type also let us evaluate the spatial relation between the origin residue and the neighbour residue. Aside from their 3D coordinates, we also stored structural information about each residue, such as their secondary structure membership. Given this information, we were able to use the atlas to generate the data in Table 3.1.

While the main purpose of the atlas was for large-scale statistical analyses, we also built a graphical user interface (GUI) that enables a user to visualize the interactions in Chimera. In the GUI, users can filter residue pairs by their types, by secondary structures, by conformers, and by their interaction energies. The GUI can be useful for visualizing clusters of similar interacting side-chains or individual samples.

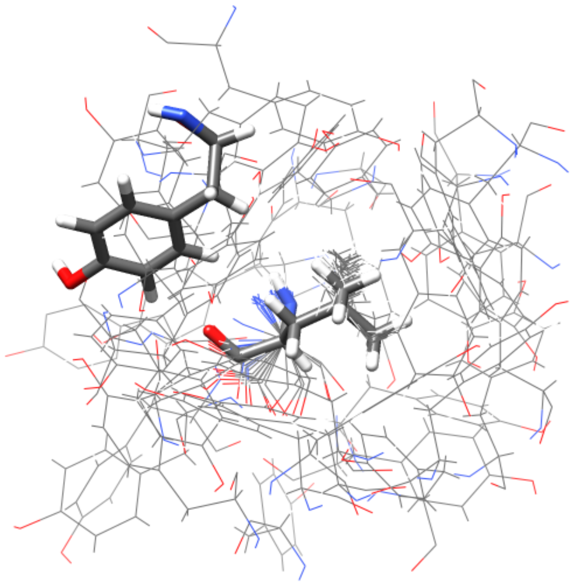


Figure A.1: Visualization of interacting leucine-tyrosine pairs using the side-chain atlas. One pair is highlighted using stick representation while all other pairs are shown with wire representation. The cluster of residues in the middle are the origin residues (leucine in this case), while the surrounding residues are the neighbour residues (leucine). Only a sample of leucine-tyrosine pairs are shown here to prevent over-cluttering the display.

Appendix B

Comparisons with Experimental Results

We verified our propagation results with functionally important residues determined from experimental studies. These experimental studies were performed by McLaughlin et al., who mutated residues and analyzed the functional cost of each mutation relative to the wild-type protein [62]. Figure B.1 shows the results of the mutational experiments that McLaughlin et al. performed.

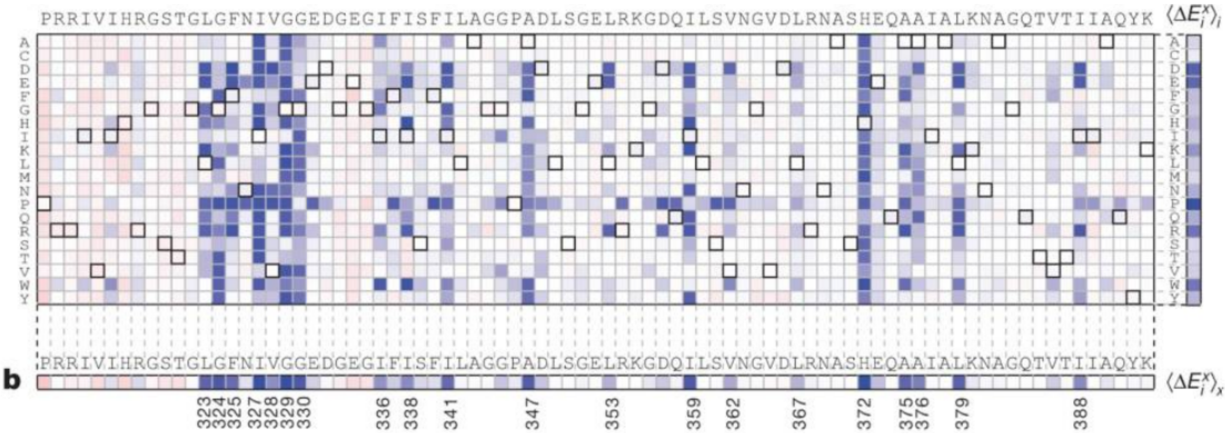


Figure B.1: Mutational analysis of PDZ3 residues with the highest functional cost residues shown at the bottom. This figure was reproduced from Figure 2 of the paper by McLaughlin et al. [62]: 10.1038/nature11500. Reproduced with permission from Springer Nature.