FUNCTIONAL DYNAMICS IN BETA-LACTAMASE: INSIGHTS INTO SUBSTRATE RECOGNITION AND INHIBITION

by

Chris Avery

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics

Charlotte

2023

Approved by:

_____
Dr. Donald Jacobs

_____
Dr. Jun-tao Guo

_____
Dr. Xiuxia Du

_____
Dr. Irina Nesmelova

_____
Dr. Jay Foley

ABSTRACT

CHRIS AVERY. Functional Dynamics in Beta-Lactamase: Insights into Substrate Recognition and Inhibition. (Under the direction of DR. DONALD JACOBS)

Beta-lactamase proteins are among the most prominent causes of antibiotic resistance. These enzymes confer resistance to beta lactam antibiotics, which are commonly used to treat bacterial infections. In recent years, novel beta-lactamase have emerged exhibiting resistance to all classes of beta lactams, representing a major threat to global health.

The mechanism by which beta-lactamase can expand their substrate specificity to confer bacteria with resistance to novel drugs is complex and not-well elucidated. In this work, beta-lactamase function is explored using a variety of computational techniques to identify the molecular mechanisms behind antibiotic resistance. In particular, the connection between protein dynamics and protein function is explored in beta-lactamase, revealing how changes in enzyme motion are related to changes in the enzyme substrate specificity.

To study beta-lactamase function, a library of molecular dynamics (MD) simulations was generated which includes simulations of TEM-1, TEM-2, TEM-10, and TEM-52 beta-lactamase, either in its apo or holo form. Holo simulations were performed with the enzymes in complex with either ampicillin, amoxicillin, cefotaxime, or ceftazidime. The enzyme-antibiotic combinations were chosen to represent both wild-type and extended-spectrum beta-lactamase activity.

To identify the functional dynamics responsible for substrate recognition, Supervised Projective Learning with Orthogonal Completeness (SPLOC) was employed. SPLOC compared the beta-lactamase simulations under different groupings to understand the role of both enzyme mutations and antibiotic interactions in determining substrate recognition. These motions were also leveraged to classify whether an

enzyme would be able to express extended-spectrum antibiotic binding. Finally, the utility of exploiting these functional dynamics to inhibit beta-lactamase function was explored using pepStream. Novel peptides were generated which bound with specificity to regions of the enzyme exhibiting functional dynamics.

This work identified dynamic signatures in beta-lactamase underlying substrate recognition. Importantly, these signatures took the form of increased flexibility in loops bordering the active site of the enzymes, which mediate local conformational flexibility that facilitates optimal substrate interactions with different antibiotics. Notably, the dynamic signatures between different protein-antibiotic systems was unique, reflecting the complexity of the mechanisms underlying antibiotic binding. A proof-of-concept for designing de-novo peptides to bind with beta-lactamase at these regions suggests that a novel class of beta-lactamase inhibitors could inhibit these motions required for substrate recognition, yielding a novel method for controlling beta-lactamase mediated antibiotic resistance.

ACKNOWLEDGEMENTS

I would like to thank all the people who have helped me throughout my time at UNC Charlotte and made this dissertation possible. First and foremost, I thank my advisor Dr Donald Jacobs, who has been the greatest influence on my scientific journey, for his unwavering support, guidance, and most importantly his patience throughout the many ups and downs that I experienced while undertaking this research. From him, I have learned about far more than just biophysics.

I would like to thank my committee members for their helpful suggestions and comments regarding my dissertation. Additionally, I would like to thank all the members of the Bio Molecular Physics Group, past and present, who have endured me talking about beta-lactamase for many years, and influenced this work through meaningful discussions during our many group meetings. In particular, I would like to thank John Patterson, Lonnie Baker, Dr Jenny Farmer, and Tyler Grear for providing invaluable insights from your respective expertises, when my own knowledge fell short.

I would like to acknowledge HPC resources provided by the UNCC Research Computing group, and give a special thanks to Jon Halter for providing advice on computational matters. This work was partially supported through the UNCC CGL's STEM Fellows Communication Program, and the Science, Mathematics, and Research for Transformation (SMART) scholarship, supported by the Department of Defense.

Finally, I would like to recognize the incredible support, both academically and otherwise, I have found in the Charlotte community. This includes so many more people than I could fit on a single page, without whom I surely would have gone insane.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Motivation

Proteins constitute one of the most crucial classes of molecules in biology, playing a pivotal role in various cellular functions essential for the existence of life. As a result, the investigation of proteins and their functions has significantly influenced research in biology, physics, chemistry, and numerous related disciplines. A full treatment of protein biophysics can be found in an introductory textbook such as [3]; To motivate the work presented here, some relevant concepts are presented in the following sections.

1.1.1 What is a Protein?

Over time, our comprehension of proteins has undergone significant advancements, shaping our fundamental understanding of their nature and functionality. Proteins are polymer chains of amino acids linked together by peptide and disulfide bonds, forming what is known as the primary structure. The sequence of amino acids that make up a protein, along with their specific arrangement, is referred to as protein sequence. There are 20 canonical amino acids which are found in protein molecules, which means for a protein whose sequence is of length N, 20^N unique protein chains can be formed. Amino acids are composed of an amino group, a carboxyl group, and a side chain residue, all connected to a central carbon atom. Amino acids can be linked together into chains through peptide bonds joining the carboxyl group of one amino acid to the amino group of another.

Amino acids can be classified based on their chemical properties. They can be either charged (positive or negative) or uncharged, hydrophobic or hydrophilic. Fur-

thermore, their side chains can exhibit various structural motifs, including aliphatic chains or aromatic rings. The biophysical characteristics of an amino acid sequence plays a crucial role in determining its interactions with the surrounding environment. In the vicinity of a cell, this environment is predominantly aqueous and encompasses a diverse array of proteins, sugars, RNA, lipids, and various metabolites that proteins can interact with. Additionally, self-interactions between amino acids within a protein can significantly impact its behavior. Intramolecular interactions, such as hydrogen bonding, are considered to be the fundamental driving forces behind the intricate process of protein folding.

Folding refers to the transformation of a protein from a disordered chain of amino acids into a well-structured protein conformation. Amino acids possess multiple hydrogen atoms, enabling the formation of hydrogen bonds. These bonds serve as nature's adhesive, allowing molecules to adhere to each other through non-covalent interactions. Hydrogen bonds facilitate the association of different segments within a protein chain, giving rise to well-organized structures like alpha helices and beta sheets. These local structures, known as secondary structures, are pervasive throughout the protein realm. Secondary structures combine to form a compact overall structurem, called a fold or tertiary structure, encompassing the entire protein chain. The process of folding a protein from a disordered chain of amino acids into a tightly-packed fold is intricate and not yet fully understood. However, it is widely assumed that the information necessary for proper folding of a protein in a given environment is encoded within the amino acid sequence, as demonstrated by Anfisen's "Thermodynamic Hypothesis" [4]. Recent advances in machine learning algorithms, such as AlphaFold2 (AF2), have demonstrated remarkable success in predicting protein structure directly from its sequence. [5]

Protein folding is governed by thermodynamics, where the most probable structure of a protein is the one that maximizes the release of energy during folding. The

energy available to a protein is quantified by the Gibbs Free Energy, defined by 1.1. In this equation, H represents enthalpy, which is the energy associated with the chemical bonds that define the protein's structure. S denotes entropy, which reflects the degeneracy of states and the degrees of freedom within the structure. Lastly, T represents the temperature of the ensemble.

$$G = H - TS. \tag{1.1}$$

In biochemistry, processes that occur in molecular systems are characterized by the associated change in free energy of the system. The likelihood of a process occurring is determined by the size of the change in free energy ΔG, where negative changes in free energy indicate processes that are more likely to occur. According to statistical mechanics, the likelihood of protein folding can be quantified by an equilibrium constant defined in Equation 1.2, where Z represents the partition function for the system.

$$K_{eq} = \frac{Z_{fold}}{Z_{unfold}} = e^{-\Delta G/k_b T} \tag{1.2}$$

The equation indicates that the probability of a protein folding is proportional to the Boltzmann factor, which accounts for the free energy change at a given temperature T. In the context of protein structure, the 'process' under consideration is the transition from a disordered structure to an ordered protein structure, involving the formation and disruption of hydrogen bonds and other non-covalent interactions.

Beyond the tertiary structure, proteins can also aggregate into multi-chain complexes, which serve as functional units and represent the quaternary structure, the highest level of protein organization. The most probable structure a protein takes for globular proteins is called the proteins native fold. It is crucial to recognize that proteins are not rigid entities but possess significant flexibility. In cellular environments, the presence of various surrounding molecules can profoundly impact the folding pro-

cess of proteins, leading to substantial reshaping effects.

1.1.2 Structure, Function, and Beyond

Protein structure is believed to be strongly linked to protein sequence and function. Notably, protein structures exhibit much less diversity compared to protein sequences. It was previously noted that most proteins folds share alpha helix and beta sheet motifs, and many sequences can share the same fold. Studies have shown that a high sequence similarity does not imply a high structure similarity, and vice versa. A cutoff at 30% sequence similarity, known as the twilight zone [6], represents the limit to where two proteins will be likely share a fold. This suggests that protein folding holds meaning beyond simple sequence, namely that protein structure is a major predictor for protein function. This relationship between protein sequence, structure and function has become a fundamental principle in protein biophysics called the sequence-structure-function paradigm. [7]

Protein function is often a loosely defined term in the literature due to its complexity, however in general it refers to biochemical and physical events involving proteins, including how proteins interact with their cellular environment. [8] A common source for different protein functions is the Gene Ontology (GO) database [9], which divides function into three categories: molecular function which includes processes such as catalytic activity or ATP binding, cellular component which describes where in a cell a protein acts, or biological process which describes what functional pathway, like DNA repair, a protein takes part in.

The complexity of protein function is as diverse as the cell needs it to be to keep itself alive. At the molecular level, even proteins that perform the same task, such as two enzymes that hydrolyze the same functional group of of a molecule, may have functional differences in respect to catalytic efficiency or substrate specificity. Changes in functionality like this most often result from point mutations that interfere with protein activity. Evolution can introduce random mutations which can

be selected for providing a gain or loss of function depending on what increases the fitness of the organism. Substrate specificity, the ability of enzymes to bind strongly with particular partners, is an important example of changes in function like this, with applications to drug design.

Together these observations regarding the relationship between sequence, structure, and function underlies the sequence-structure-function paradigm. The sequence determines how a protein folds into a structure, which has been evolved to facilitate a particular function. Mutations occur independently to change the sequence, subtly specializing the structure and function of the protein. Over time, numerous mutations can accumulate, leading to a family of proteins with related structures, sequences, and functions. Over long time periods, divergent functions may arise, however in general, mutations which become fixed on a protein often cause small shifts in function that may benefit its parent organism due to some environmental adaptation. The underlying idea is that when a structure has been evolved to perform a function, it tends to undergo small, rather than large, functional changes.

Sequence, structure, and function are the foundation for understanding how proteins work. The development of both experimental and computational methods have been instrumental for giving researchers a tangible glimpse into the world of proteins. These advancements have enabled the elucidation of function in many proteins at the atomic level, allowing for the study of biochemical mechanisms of complex reactions. Despite these strides, how to characterize or alter protein function still remains complex, with the exact role of point mutations and the environment on function not being fully understood and difficult to model and predict.

An area rapidly gaining interest in the wider protein science community is the role of protein dynamics in determining protein function. [10, 11] Protein dynamics refers to the movement of proteins in their environment due to thermal fluctuations and external stimuli. Proteins in their native state can be rigid or flexible. Some proteins,

called intrinsically disordered proteins (IDP), do not settle into an ordered structure at all and are characterized as having high flexibility. [12] Often proteins do not have just one stable conformation within its fold, but can transition between a few different conformations as it executes its function. [13] Even slight changes in conformation can change the function of a protein. A reorientation of secondary structure or just a few residues can disrupt the complex network of interactions between a protein and its ligand, leading to a gain or loss of function. Mechanisms like allostery use global and local conformational change to regulate the function of a protein. [14] Consequently, the role of dynamics in protein function has become prominent in protein biophysics and protein engineering research, and will be the subject of this work.

1.2 Problem Statement

Proteins play an important role in defending cells from harmful environmental invaders. In bacteria, these proteins can be used to confer antimicrobial resistance, allowing bacteria to evade drugs that would otherwise kill them.

Antimicrobial resistance (AMR) poses a significant threat to public health. According to the Antibiotic Resistance Report in 2019 from the Center for Disease Control, there are over 2.8 million AMR infections per year in the US. [15] Antibiotics have been a primary weapon against bacterial infections for over 70 years, however widespread AMR poses a major threat to their continued efficacy. The emergence of AMR towards a new drug can develop when bacteria are exposed to a new medicine. A small fraction of the bacteria may survive due to random mutations enabling the surviving bacteria to pass on their preferred traits to future generations, resulting in resistant strains. [16, 17] In response, measures have been put in place to reduce the spread of AMR, such as measures to prevent antibiotics from being released into the environment and to control the number of extraneous prescriptions. [18, 19] However, novel approaches for combating AMR microbes are needed.

One approach to address AMR is to engineer novel drugs that can circumvent a

pathogen's defenses. Achieving this requires a detailed understanding of antimicrobial resistance mechanisms and the complex process of drug design. Drug design is a complex process which uses High Throughput Screening (HTS) to search for potential molecules that have a high affinity for a target, and highly parallel affinity experiments to test them. [20, 21] However, this method is time-consuming, costly, and high-risk for each new drug approved for use. Additionally, HTS often only provides a limited understanding of the mechanisms behind how drugs work. To address this computational approaches have been implemented including both ligand-based and structure-based approaches which leverage large databases, structural data, and more recently machine learning.

Over thousands of years bacteria have developed many antimicrobial resistance mechanisms, and one of the most common is a family of proteins called beta-lactamase. Many thorough reviews on beta-lactamase-based AMR exist, such as [22, 23], and some of the relevant information is presented here. Beta-lactamase confer resistance against a class of antibiotics called beta-lactams, which includes penicillins, cephalosporins, monobactams, and carbapenems. All beta lactams share a common structural feature called a beta lactam ring, a 4-member ring motif. Their mechanism of action is to bind to a cell wall transpeptidase (penicillin binding protein), preventing the cross-linking of peptidoglycans which hold the bacterial cell wall together, ultimately leading to cell death. Beta-lactamase are hydrolases, thought to be directly descended from transpeptidases [24], that intercept the drugs and catalyze a reaction that breaks a bond in the beta lactam ring which is critical for recognition and binding to transpeptidase proteins.

The super-family of beta-lactamase proteins is very large, encompassing thousands of enzymes. Among them, the most commonly encountered are TEM, SHV, CTX, and OXA families. TEM-1, discovered in 1963, is considered the wild type enzyme as it was initially the most prevalent. During the "golden age" of drug discovery

during the 1950's and 1960's many new antibiotics were rapidly developed which led to a significant evolutionary pressure on bacteria to adapt beta-lactamase enzymes for better survival, which in turn led to the explosion of beta-lactamase diversity that is now seen today. A crucial factor in this diversity is that beta-lactamases are plasmid encoded genes, which allows for horizontal gene-transfer between bacterial communities.

While all beta-lactamase share the same catalytic function, each enzyme is fine tuned to exhibit specificity for certain antibiotics. For example, TEM-1 primarily acts on penicillins, while CTX enzymes prefer cephalosporins, and OXA enzymes favor oxacillin. This subtle change in function between enzymes poses a significant challenge to public health in combating beta-lactamase-mediated resistance. Although beta-lactamase enzymes show significant sequence diversity between families, changes in substrate specificity can also vary greatly within a family where sequence is highly conserved, as is the case for the TEM family. This adaptability comes from point mutations which influence enzymes specificity through structural and dynamic changes, although the exact mechanisms that control substrate recognition are unknown. These changes, in tandem with antibiotic misuse, has created a feedback loop that amplifies the emergence of novel beta-lactamase enzymes, ultimately leading to the current rise in extended-spectrum, carbapenem, and inhibitor resistant beta lactamase around the world (ESBLs, CBLs, ard IRBLs, respectively).

1.3 Research Objectives

This project focuses on understanding changes in beta-lactamase function, specifically substrate specificity, in response to mutations to its sequence. Important structural and dynamic changes in the enzyme are identified using a variety of bioinformatic tools. Prior to presenting the results, in Chapter 2 a computational biology background to beta-lactamase-mediated AMR will be given along with a discussion of current methods in computational studies of protein function and dynamics. Chapter

3 will introduce the main tools developed or utilized in this project.

This project addresses two scientific questions: 1) Can the dynamic impact of mutations on beta-lactamase be discerned to elucidate how the enzyme develops resistance to different antibiotics? 2) Can this information be applied to inform the computational drug design process. As part of Aim 1, Chapter 4 uses traditional and new techniques for analyzing molecular dynamics (MD) simulations of beta-lactamase. This reveals the importance of dynamic allostery, and quantifies the functional dynamics which drives beta-lactamase function. Aim 2 explores the efficacy of using peptides as beta-lactamase inhibitors to combat AMR. In Chapter 5 pepStream, a method of predicting *de-novo* peptide binders for specific regions on a proteins, is employed to predict binding peptides which target regions on beta-lactamase that express functional dynamics as determined in Chapter 4. Finally in Chapter 6, the conclusions of these studies are presented and summarized and discussed in a broader context.

CHAPTER 2: BACKGROUND

In this chapter necessary biophysical and biochemical background on beta-lactamase is presented.

2.1 Functional Dynamics in Proteins

Recognizing the link between protein dynamics and protein function is now widely acknowledged; however, identifying functional motions poses a significant challenge. Proteins exhibit motion on multiple scales of time and structure. The largest, and often slowest, motions are associated with folding, where a chain of amino acids transitions from random coil to compact domain. At the next level down are domain motions which control conformational changes within a protein, both in terms of global changes to the whole domain and local changes to particular structure elements. The smallest motions are sidechain dynamics, which include atomic motion such as rotations between side chain rotamers down to bond vibrations. Each of these motions plays a crucial role in determining how a protein functions.

Protein dynamics has remained elusive because they are challenging to directly observe. X-ray crystallography, the most prominent structure determination method (representing 85.3% of the Protein Data Bank (PDB) [25] structures as of July 5, 2023), only provides a single structure. In contrast, nuclear magnetic resonance (NMR), the third most prominent method (representing 6.77% of the PDB structures as of July 5, 2023), offers multiple conformations, enabling a glimpse at the potential conformational dynamics of proteins. B-factors from crystallography have been observed to loosely correlate with protein flexibility and has served as an indicator of protein motion. [26] Infrared spectroscopy can be used to probe protein

dynamics as domain-level molecular motions are detectable at these wavelengths. [27] However, this information may be insufficient to provide atomistic details about molecular mechanics.

To bridge this gap, many studies rely on computational modelling of protein dynamics in combination with these experimental methods. Molecular force fields have been developed based on theories of intermolecular interactions that are calibrated against experimental data. While the models are far from perfect, they have been shown to reasonably approximate the physics of proteins in solution, and have proven useful in aiding experimental studies. [28]

The primary drawback of using computational models to study protein dynamics is the vast number of timescales that functional protein motions can occur. This can range from the picoseconds to hours or days, while MD simulations are typically limited to the micro-millisecond range. Additionally, for large proteins it becomes nearly impossible to simulate motions at this scale due to hardware constraints. Sufficient sampling is needed to fully understand molecular processes, thus these challenges can be addressed to some degree by using coarse grained MD models [29], accelerated molecular dynamics, or Markov state models. [30]

Ultimately, the goal of thorough sampling is to characterize the conformational free energy landscape (FEL) of a protein. Proteins can exist in a multitude of microstates represented by local minima of the FEL. [11] For example, two microstates could represent the starting and ending conformation for a protein undergoing a specific process. A transition between one state and another represents the process proceeding. The functional dynamics of the protein are often those which allow for the protein to move between microstates.

The free energy change (ΔG) associated with such processes is related to relative depths of the microstate minima on the FEL. Obtaining accurate ΔG values would require sufficiently sample every possible conformation available to the protein. Due

to the nature of proteins this is infeasible, so concessions are often made to adequately sample just a portion of the FEL relevant to the problem.

2.1.1 Flexibility and Function

In the thermodynamic context, globular proteins have a native state, which corresponds to their folded and functional conformation. This state exists at the point of lowest free energy on the FEL, resembling a funnel or basin. Current research indicates that, locally, the FEL is actually rugged, giving rise to various partially folded microstates the protein can access as it transitions from random coil to folded. [31, 32] Additionally, the bottom of the basin can also be rugged and broad. In the case of a rugged, flat-bottomed basin, a degeneracy of local minima within the flat bottom, manifests physically as multiple equally likely folded conformations which the protein can transition between in equilibrium, similar to the example given above.

Transitions between conformation states are facilitated Brownian motion. Thermal fluctuations within the protein and its surroundings cause continuous small fluctuations around the native state structure. If a sufficiently large random fluctuation occurs, the protein can be displaced from its current conformation enough to allow it to enter another microstate. The propensity for a protein to escape its current basin is controlled by the free energy barriers between states, and this process can either be driven entirely by random thermal motion, or by external stimuli.

To be functional, a change in conformation must result in a modification in the protein's ability to perform its specific task. For example, a protein may fold in such a way that its active site is initially obscured from the solvent-accessible surface, so that interaction with a ligand is unfavorable. After conformational change, however, the active site is opened and interactions are favorable. In this case the change results in the protein being able to interact with its ligand. Often times the environment plays a crucial role in triggering functional changes in proteins. This is one of the many ways that the molecular environment is related to protein function.

In the case where random fluctuations are not enough to completely eject a protein from its local minimum, it tends to return to equilibrium and reassert its native conformation. The extent of this elasticity in proteins, to deform its shape, is referred to as the native state flexibility of the protein. Flexibility can manifest at all levels of the proteins structure, including its global domain shape, secondary structures, and even down to side chain mobility.

Proteins can leverage flexibility in a variety of functional ways. For example, a protein can reposition amino acid side chains for optimal interacting poses or contribute to maintaining the structural integrity of a fold. [33] Flexibility can also arise by random chance through brownian motion as described above. Discerning flexibility which arises from random chance versus as part of protein function is a challenging task.

2.1.2 Detecting Functional Flexibility

Consider a set of sampled conformations of a protein, $C = \{c_i\}$. Each c_i represents a vector of x, y, and z coordinates of each atom in the protein, $c_i =< x_1, x_2, x_3, ..., y_1, y_2, y_3, ..., z_1, z_2, z_3 >$. The index i represents the ith time step of the trajectory. To assess flexibility in proteins, some simple metrics can be employed.

To allow meaningful comparisons between changes in atomic positions throughout a trajectory, a base line reference is required. [34] The true native state structure would be ideal, however this is not attainable. Instead, the initial conformation of the trajectory x_0 or the conformation closest to the average can be used in its place. During this discussion the reference will be called x_{ref}.

The Root Mean Square Deviation (RMSD) [35] is a commonly used metric that is for assessing the change in a proteins structure throughout a trajectory. For any single conformation, the RMSD is calculated the average deviation of all atoms compared to the reference as in Equation 2.1.

$$RMSD(x_i, x_{ref}) = \sqrt{\langle (x_{ref} - x_i)^2 \rangle_{atoms}} \tag{2.1}$$

In the above equation $\langle \cdot \rangle$ represents an average over the atoms per conformation. The RMSD provides a single numerical value to compare to the average deviation of a protein from a reference.

To gain a more detailed understanding of where flexibility along a protein chain occurs, Root Mean Square Fluctuation (RMSF) can be used. This metric averages the deviations between each conformation and a reference per atom basis over the full trajectory. The RMSF provides insights into the fluctuations and variations exhibited by individual atoms across the set of conformations.

$$RMSF(\{x_i\}, x_{ref}) = \sqrt{\langle (x_{ref} - x_i)^2 \rangle_i} \tag{2.2}$$

RMSD and RMSF should be carefully considered to avoid arriving at misleading conclusions. As an example, a large value for RMSD may actually indicate that the trajectory has moved away from the reference conformation but is rigidly sampling this divergent pose. This would raise the value without increasing actual flexibility. Additionally, a common reference structure when comparing values across different proteins so that all values can be compared on equal footing.

RMSD and RMSF provide some insight into the motions atoms in a protein can exhibit beyond a single structure. However, protein motions are often more complex and function can involve cooperative motion across the protein. Furthermore, it is difficult to distinguish what motion is functional versus non-functional. For this, more complex models have been developed. Recently, machine learning has been exploited for understanding protein dynamics. [36] These advanced methods provide a deeper understanding of the complex motion involved in protein function.

2.1.3 Internal Dynamics of Proteins

Global metrics of flexibility, such as RMSD and RMSF, provide insights into protein behavior, however they can fall short in capturing the full complexity of protein function. The internal motions of proteins, such as conformational changes or hinge bending motions, can also facilitate various biological processes involving proteins. However, identifying and modeling these dynamics pose significant challenges. Nonetheless, gaining a comprehensive understanding of protein dynamics provides a detailed picture of how proteins interact with their environment, revealing the underlying mechanisms of their functionality.

2.1.3.1 Elastic Network Models

A common approach to detecting functional motions in proteins is to build an Elastic Network Model (ENM), which are extensively reviewed in [37]. ENMs characterize the global vibrational motions of a protein by treating proteins as a network of springs and computing the normal modes of the network. The lowest frequency motions correlate to the largest amplitude structural motions of the protein.

ENMs can be constructed from a single protein structure. There are several types of commonly used ENMs for protein analysis, including Gaussian Network Models (GNM) and Anisotropic Network Models (ANM). GNM models fluctuations about the equilibrium state of the protein without considering the directionality of the motion, while ANMs include direction information. In GNMs, the spring network is constructed using distances between atoms and their equilibrium positions, resulting in an n dimensional network for an n-atom protein. In ANMs displacements in all $3n$ directions are considered, resulting in a $3n$-dimensional network. The following discussion will focus on ANMs.

ENMs can be used to perform fast, coarse-grained simulations of protein motion, serving as an alternative to computationally expensive MD simulations. [38] They

have been found to perform well in reconstructing conformational changes in proteins and in capturing important fluctuations as described by b-factors in x-ray crystallography. [39]

In an ENM, atoms in a protein structure are represented as masses connected by springs, representing bond constraints. Consequently, the whole system can be thought of as a set of coupled oscillators. The equilibrium structure, denoted as $r^0 = [x_i^0, y_i^0, z_i^0]$ where i is an index over all atoms, represents the configuration of oscillators at equilibrium. This system can be treated in a way similar to a network of springs as can be found in any introductory Classical Mechanics textbook, such as [40]. Expanding around equilibrium, the potential energy of the system can be expressed up to the second order as:

$$V(r - r^0) = V(r^0) + \sum_i (\frac{\partial V(r^0)}{\partial r_i})(r_i - r_i^0) + \sum_{i,j} \frac{1}{2}(\frac{\partial^2 V(r^0)}{\partial r_i \partial r_j})(r_i - r_i^0)(r_j - r_j^0) \quad (2.3)$$

The first and second term are zero by definition, as the system being in equilibrium. The term $\frac{\partial^2 V(r^0)}{\partial r_i \partial r_j}$ defines the elements of the Hessian matrix H. The fluctuations about the equilibrium position can be represented by the $3N$ dimensional vector $\Delta r = [(x - x^0)_i, (y - y^0)_i, (z - z^0)_i]$, which simplifies the potential energy for small fluctuations about the equilibrium state to 2.4.

$$V(r - r^0) = \frac{1}{2} H_{ij} \Delta r_i \Delta r_j = \frac{1}{2} \Delta r^T H \Delta r \quad (2.4)$$

When this is substituted into the Euler-Legrange equations, the equations of motions for the system can be obtained as in 2.5.

$$M \Delta \ddot{r} = -H \Delta r \quad (2.5)$$

This takes the form of Hooke's law describing harmonic motion. The solutions for this is well-known and takes the form of a set of n modes with the form 2.6.

$$\Delta r_i = A_i cos(\omega_i t + \phi_i) \tag{2.6}$$

When this is substituted back into the equations of motion, with some algebra, the vector A_i becomes the solution to an eigenvector equation shown in 2.7.

$$M\omega_i^2 A_i = H A_i \tag{2.7}$$

The operator that needs to be diagonalized to obtain the normal modes is $K = M^{-1}H$, known as the mass-weighted Hessian. The eigenvalues of this operator represent the squared frequencies of oscillation for the corresponding normal mode. It is important to note that the eigenvalues do not have to be positive by definition. A positive eigenvalue indicates the frequency of a harmonic oscillation in a local minima of the free energy landscape, while a negative eigenvalue leads to an imaginary frequency which represents saddle points in the free energy landscape or transition states of the molecule.

This model for describing the internal motions of proteins is directly motivated from fundamental physics principles. Correctly parameterizing the matrix H (or K) is crucial for capturing the true physical motions of the system. In most models, the form of this matrix is determined by the connectivity of the of the spring network, as illustrated in Equation 2.5. The value of H_{ij} is equal to the spring constant between atoms i and j in the network, and correspondingly, for two non-interacting residues this value will be zero as they do not exert force on each other. For interacting residues it will be positive and correlate to the strength of their interaction.

For a GNM, the H matrix is defined to be kI, where k is a parameter of the model and I is the identity matrix. On the other hand, ANMs have more sophisticated

structure because they separate the three components of fluctuation vectors, thereby providing more information about the normal modes of the protein. [37] In general, the exact definition of the H_{ij} is not explicitly defined, however it is common to define them such that closer atom pairs have higher spring constants. Additionally, a common practice is to use a cutoff distance for considering two atoms as interact with each other, where H_{ij} is 0 for atoms further than this cutoff.

ENMs rely on several assumptions which must be considered when evaluating the validity of the model. Firstly, ENMs are valid in the linear response regime for the network, which practically means that it can only describe atomic motions that are small relative to the natural length of the springs connecting them. The concern is that the force constants may change dynamically as the system evolves. For example, consider the case that two residues start near each other and therefore have a non-zero k value. As the protein changes its structure, which would break the assumption, the atoms move apart resulting in a decrease in the k value over time.

Another consideration is that ENMs are typically constructed using a single protein structure, which may overlook important environmental effects which influence the protein dynamics. Solvent molecules and other small molecules are typically not represented in these models. Finally, normal mode motions are still approximations to the actual dynamics a protein might undergo in reality. The normal modes can only describe motions up to the level of coarse graining that was used in the underlying elastic network describes.

For fluctuations around the equilibrium structure in the quasi-harmonic approximation, the Hessian matrix has been shown to be approximately equivalent to a precision matrix, or the inverse of a covariance matrix derived from the frames of an MD simulation, $H \approx kT\langle Q^{-1}\rangle$. [41] Given this observation, MD simulations can be used to estimate an ENM for a protein. This approach removes the need to parameterize a Hessian matrix as the elements originate directly from sampled protein

motion. No simulation can sample all possible conformations a protein can take, and so the resulting Hessian will also only be approximate. Since the eigenvectors for the covariance matrix are identical to those of its inverse (with inverted eigenvalues), finding the eigenvectors of the covariance matrix, a technique referred to as Principal Component Analysis (PCA), is equivalent to finding the normal modes of motion using an ENM. This approach is called Essential Dynamics (ED).

2.1.3.2 Essential Dynamics

Essential Dynamics aims to find the largest amplitude motions of a protein. [42, 34] In the context of ENMs, these high-amplitude motions correspond to the lowest frequency normal modes. [43] To perform ED, a trajectory of protein conformations can be sampled with MD or experimental methods. Prior to analyzing the dynamics, all frames of the trajectory must be aligned using structural superposition to remove trivial degrees of freedom describing global translation and rotation.

PCA is the method which underlies ED. The primary objective is to transformation the original data into a set of collective variables which maximize variance. [44, 45] A collective variable, Equation 2.8, is a generalized coordinate of a system which describes a global feature of the original data. [46, 47]

$$q(t) = \Sigma_i c_i x_i(t) = \langle v_i | x(t) \rangle \tag{2.8}$$

For a protein, the vector of atomic positions at time t serve as the observables of the system, denoted by $x_i(t)$. Generally, a collective variable is a linear combination of the observables, weighted by a set of coefficients $\{c_i\}$. Learning this set of coefficients which maximizes the variance of $q(t)$ is the goal of PCA. Mathematically, the desired coordinates correspond to the eigenvectors of the covariance matrix, where the value of each c_i is defined by the components of each eigenvector, $|v_k\rangle$. These eigenvectors can be computed by solving the following eigenvalue problem. The eigenvalue

corresponding to $|v_k\rangle$ is λ_k.

$$Q\,|v_k\rangle = \lambda_k\,|v_k\rangle \tag{2.9}$$

To find the most variant motions of the protein, the matrix Q in Equation 2.9 is equal to the covariance matrix computed from the trajectory. In this context, the eigenvalues represents to the variance of the motion that is described by the associated eigenvector. A collective variable, called the principal component, can be obtained projecting the observations onto an eigenvector.

The "essential dynamics" of the protein correspond to the motions described by the eigenvectors with the largest eigenvalues. Under the quasi-harmonic approximation, the essential dynamics are equivalent to the normal modes with the lowest frequency. This correspondence between essential dynamics and low-frequency motion explains why the lowest frequency motions are often considered to be the motions with the largest amplitude.

For a protein with $3N$ total degrees of freedom (N being the number of atoms in the structure), the number of principal components needed is highly dependent on the system and dynamics being observed. However, the choice of how many components to keep is often left to more quantitative metrics such as requiring a minimum percent of the original variance or using Cattell's criterion. [48]

The choice of matrix Q is not limited to the covariance matrix of Cartesian coordinates, however, only the covariance matrix has the special interpretation of representing low-frequency vibrations. In principle, any symmetric and positive definite matrix of observables from an MD trajectory can be used to describe protein motion. For a matrix which satisfies these conditions, the eigenvectors will form complete orthonormal basis set, valid for constructing collective variables to describe features of the system. The motions described by different matrices, and constructed from different coordinate types in PCA will reflect different assumptions and have different

interpretations. [49]

A common matrix that can be used is the correlation matrix, R, which describes the pairwise correlation between observables, $R_{ij} = \rho(x_i, x_j)$. [49] In correlation-based, the eigenvalues represent the proportion of original variate correlation is represented in each principal component. A rule of thumb for correlation PCA is that an eigenvalue greater than 1 indicates that the eigenvector encompasses at least one variable's worth of correlation.

PCA is a useful method for identifying functional motions in proteins due to its ability to automatically identify the most variant or correlated dynamics of the system. When treated as a linear transformation into a one-dimensional collective variable, each eigenvector provides an interpretable model for identifying which atoms contribute to the motion it represents.

PCA belongs to a class of algorithms called projection pursuit or dimensionality reduction, which stems form the algorithm of the same name [50], in which high dimensional data is projected into a low dimensional representation. [51] In PCA, dimension reduction is done by dropping the eigenvectors which are below a threshold for variance or correlation. This filtering process eliminates motions that are typically considered random fluctuations or noise. PCA can also be applied to a variety of different types of degrees of freedom, including dihedral angles and atom to atom distance pairs, in order to identify different types of motions. Generally, PCA is well suited for identifying large scale motions like conformational change or hinge bending, e.g. in DNA binding protein domains. [52]

ED is usually applied to just the alpha carbon atoms in a protein, as early studies demonstrated that coarse graining a protein to just the backbone or alpha carbon atoms is sufficient for identifying large motions in proteins. [53] Higher resolution models are tend to be noisy, as the large-scale backbone motions wash out the information from smaller side chain motions. [54] The primary benefit of using coarse-grained

models is that the reduction in computational burden associated with computing essential dynamics. However, ED can be effectively applied to a small subset of atoms or residues at the all-atom or heavy atom level, allowing for an investigation into the role of side chain motions in protein function. This type of funnel approach can provide valuable insights into the role protein dynamics play in function, however it usually requires expert knowledge about the system to curate such a set of atoms.

2.1.4 Allostery

In the context of functional flexibility, protein function is driven by the conformational capacity of the protein itself. However, in the case of allostery, function is driven by ligand binding. Ligand binding is a fundamental mechanism used by proteins to perform functional processes, whether this is catalysing reactions or transmitting inter-cellular signals. Allostery is a phenomenon that occurs in proteins with two or more distant binding sites, either on the same or separate chains. [14, 55] It involves a binding event at one binding site influencing the affinity for binding a ligand at another site, often the active site. The change in affinity can be quantified by a change in $\Delta G_{binding}$, which is called $\Delta\Delta G$. Negative $\Delta\Delta G$ indicates positive allostery, where binding one ligand increases the affinity for binding another. Conversely, positive $\Delta\Delta G$ is negative allostery, where binding one ligand decreases the affinity for binding another.

The molecules that induce allostery in proteins are called allosteric effectors. When both allosteric effectors are the same kind of molecule they are called homotropic, otherwise, they are heterotropic allosteric effectors. Allostery arises because events at distant sites on a protein can still influence each other through a mechanism called cooperativity. Allosteric signals can be transmitted between binding sites via allosteric pathways, which are mediated through conformational change and perturbations in the normal mode motions of a protein. [56] Although allostery was once thought to be a rare phenomenon, more recently it has been recognized that allostery may be a

universal mechanism in all proteins [57].

Hemoglobin, which is responsible transporting oxygen around a body, is one of the most well-known examples of allostery, reviewed in [58]. Hemoglobin is comprised of four subunits, each capable of reversibly binding with a single oxygen molecule. Studies of the Hemoglobin binding curve implied that cooperativity between oxygen binding sites was present, particularly that binding an oxygen on one subunit increased the affinity for binding on the other subunits. Further study of this system revealed a structural basis for the observed cooperative effect. When an oxygen binds into a subunit, an allosteric signal is propagated across the quaternary structure of hemoglobin through interfacial contacts between monomers. This signal induces a conformational change in the binding site that improves the affinity of oxygen binding, leading to the observed cooperative effect.

Hemoglobin is an example of structurally transmitted allostery, however, allostery can occur through two main mechanisms: conformational and dynamic allostery. Conformational allostery occurs when the first binding event triggers a conformational change at another site, modulating the affinity for a binding event to occur at this site. This mechanism is reminiscent of the induced fit or conformational selection model [59] for ligand binding which states a ligand can only bind to a protein when both molecules adopt mutually favorable conformations. In comparison to allostery, the change in binding site conformation occurs when both receptor and ligand are present in appropriate conditions. Comparatively, in allostery the conformational change must be catalyzed by effector binding, regardless if the other conditions are met.

Classic models of conformational allostery treat it as a two-state system with the two conformational states called tensed (T) and relaxed (R). A molecule in the T state has a lower affinity for ligand binding, while in the R state it has higher affinity for ligand binding. In the 1960's, the Monod-Wyman-Changeux (MWC) model [60] was

proposed to explain allostery. In this model the T and R state were pre-existing in the conformational ensemble of the protein in equilibrium, defined by an equilibrium constant. In the MWC model, the binding constants of the ligand to the T and R states are the same at each binding site, and independent of the state of any other binding sites. A binding event at any site triggers a shift in the conformation equilibrium. With each binding (or unbinding) event the equilibrium changes more and all the sub-units become more likely to change from T to R (or vice-versa). This called the continuous model.

Shortly after the MWC model was introduced, the Koshland-Nemethy-Filmer (KNF) model, or sequential model, was proposed. [61] In the KNF model, the change in conformational equilibrium only impacts binding subunits which were interfacially linked. The sequential model suggests that binding affinity across the whole protein changes as a chain reaction. These models are still used today, although, it has been shown that a protein may have more than one T or R state. The models have been updated to reflect this reality.

The MWC and KNF models have been successful in explaining allosteric mechanisms in proteins, however the models still contain some limiting assumptions. Firstly, these models addressed allostery that is present in proteins composed of identical (symmetric) binding monomers. In further work such as in [62], cooperativity between ligand binding events were explored from the perspective of separate contributions from conformational changes to the ligand binding domain and quaternary structure, as in the induced fit model. Secondly, it was assumed that conformational change was only mechanism which could cause allostery.

In 1984 Cooper and Dryden introduced a new model that did not require specific conformational change mechanisms. [63] They provided a thermodynamic proof which quantified $\Delta\Delta G$ due to ligand binding in terms of a contribution from conformational change as well as changes to the internal motions of proteins. This model proposed

that a sufficiently large change in binding affinity can come from a shift in the normal modes when an allosteric effector bound to the protein. Because normal modes are collective motions, changes in the proteins motion could be easily transmitted around the protein and between distal binding sites.

To illustrate their concept, the authors consider a situation involving a protein which has two identical binding sites, which can be occupied in three states: apo E, singly occupied EL, or fully occupied EL_2. The change in affinity between singly occupied and fully occupied can be approximated by the shift in normal mode frequencies, as described by Equation 2.10.

$$\Delta\Delta G = -kTln(\frac{\nu_1^2}{\nu_0\nu_2}) \tag{2.10}$$

Here, ν_i represents the frequency of normal modes for the states with $i = 0, 1, 2$ occupied binding sites.

2.2 Beta Lactams and Beta-Lactamase

Antibiotics are the first line of defense against bacterial infection. In 2021, the CDC reported 210.9 million antibiotic prescriptions in the United States alone. [64] Among these prescriptions, the top two classes of drugs were beta lactams, specifically penicillins and cephalosporins, with a combined total of 77.8 million prescriptions. Additionally, an additional category known as "beta lactams with increased activity" contributed an extra 22 million prescriptions. The beta lactam amoxicillin, a derivative of penicillin, is the most commonly used antibiotic worldwide. Beta lactams are an important class of drugs used worldwide, and so safeguarding their efficacy is a primary goal for maintaining global health.

Beta lactams refer to several classes of drugs which share a common chemical motif called a beta lactam ring. [23] The 4-membered ring is characterized by a square structure comprised of three carbon atoms, one nitrogen atom, and an oxygen atom

connected to the carbon of the C-N bond. The four major classes of beta lactam antibiotics are penicillins, cephalosporins, monobactams, and carbapenems. Penicillins, being the oldest and first clinically approved class of beta lactam, was discovered serendipitously by Alexander Fleming in 1928 [65] when he reported a petri dish of bacteria where some mold had contaminated the sample and stopped the bacteria cultures from growing. From this benzylpenicillin, or Penicillin G, was developed and first clinically deployed by the early 1940s. [66] Today, a variety of common penicillins are still widely used including amoxicillin, ampicillin, and piperacillin. The core structure of penicillin-like drugs is the beta lactam ring with a thiazolidine ring attached on one side and an interchangeable residue on the other side. The identity of the residue determines the identity of the compound. Other classes of beta lactam preserve the beta lactam ring but have different structures.

Cephalosporins have similar core structures to penicillin. However, in cephalosporins, the 5-member thiazolidine ring is replaced with a 6-membered dihydrothiazine ring. This change allows for a second interchangeable residue group and increases the general size and weight of the drugs compared to penicillin. Cephalosporins were first found in the 1950's and have become a vital group of antibiotics. [67] They are divided into five generations which are grouped based on their spectrum of activity. Later generations include activity against a broader range of bacteria including both gram-positive and gram-negative bacteria.

Monobactams have a more significant shift in structure, lacking a second ring fused to the beta lactam ring. Instead they consist of just the beta lactam ring with two residue groups. Azteronem is the most common example of monobactam antibiotics. Finally, carbapenems have historically been the most clinically effective class of beta lactams. They have been able to evade common forms of resistance to beta lactam drugs because they are extremely structurally stable and have high affinities for their target. Carbapenems have the central beta lactam ring which is fused to a 5-

membered carbon ring, with various additional residues. Because carbapenems have a wide spectrum of activity and have been effective for treating multi-drug resistant bacterial infections, they are some of the most important drugs currently in use.

All classes of beta lactam antibiotics generally share the same mechanism of action on bacteria. [68, 69] Bacteria often posses a rigid cell wall to maintain cellular structure and add an extra layer of defense against harmful environmental toxins. Bacterial cell walls consist of peptidoglycans, which are tightly cross linked into a mesh-like structure by transpeptidases. These transpeptidases responsible for cross-linking peptidoglycans via a short peptides are the primary target of beta lactams and as such are also called Penicillin Binding Proteins. It was observed that penicillin molecules irreversibly bind with a Serine residue in the transpeptidase active site, thereby interrupting cell wall synthesis. Consequently, the cell wall is destabilized, eventually leading to bacterial death. Since eukaryotes lack cell walls while bacteria possess them, cell walls have become a significant target for antibiotics.

2.2.1 Beta Lactamase Structure and Function

Bacteria can acquire resistance to beta lactams in three main ways. [70] First, modifications to the PBP targets can make binding with beta lactams much less favorable. Second, the use of efflux pumps controls the concentrations of antibiotics near the cell wall. Finally, and most commonly, bacteria can produce a protein known as beta-lactamase. Beta-lactamase is a large superfamily of proteins that share a similar biochemical function: catalyzing a hydrolysis reaction with beta lactam antibiotics, leading to the cleavage of the C-N bond necessary for PBP binding. This binding event is quickly turned over, allowing beta-lactamase to rapidly reset its active site and find another antibiotic to deactivate.

The beta-lactamase superfamily can be divided into structural categories A, B, C, and D (Ambler classification [71]) or into functional categories based on their substrate spectra (Bush-Jacobi classification [72, 73]). Classes A, C, and D share a

common Serine-based mechanism of action. A conserved serine residue is activated and attacks the carbonyl atom on the beta-lactam ring, forming a covalently bonded acyl-enzyme complex. A water molecule is then activated and cleaves this bond, leaving the beta lactam ring broken. When the ring is broken, the molecule is unable to bind to the PBP target, rendering the drug inactive. Serine beta-lactamases are thought to have evolved directly from PBPs on bacterial cell walls as both active sites contain a conserved SXXK motif for substrate recognition. [24] Class B beta-lactamase, which are metalloenzymes, employ a zinc mediated enzymatic pathway for inactivating beta-lactam molecules. [74] These metalloenzymes are of particular concern for public health as they can evade traditional beta-lactamase inhibition mechanisms designed for serine based enzymes. Beta-lactamase have long been of interest in biochemistry due to their clinical significance and their status as diffusion-limited enzymes with fast hydrolysis speeds. [75]

Classes A, C and D beta-lactamase are structurally conserved families, despite having significant sequence diversity. The structure of these beta-lactamase consists of a catalytic alpha-beta-alpha domain with the active site pocket bordering a beta-sheet wall, with an all-alpha terminal domain, shown in Figure 2.1. An important omega loop found adjacent to the active site in serine beta-lactamases, has been has been shown to play a crucial role in functional mechanisms of the enzyme. [76] NMR [77, 78] has shown that this loop exhibits high flexibility on longer (ms and beyond) timescales, and markov state models have been used to predict the opening of a cryptic pocket near this loop. [79]

Experiments where the omega loop was removed found a significant decrease in beta-lactamase activity [80], attributed to structural changes and the loss of catalytic residue E166. Given its proximity to the active site, several residues on the omega loop, such as E166 and N170, are proposed to directly contribute to antibiotic hydrolysis. Competing theories regarding the mechanism for activating the catalytic water

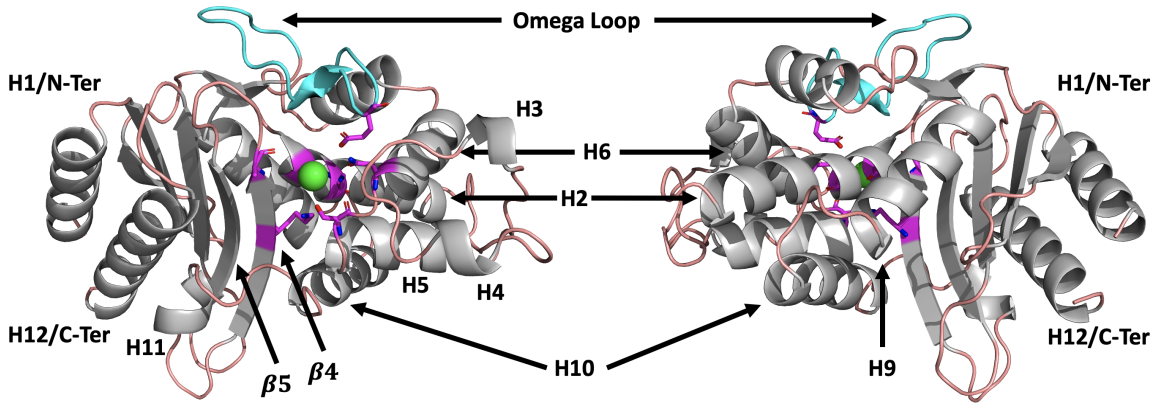

Figure 2.1: Structure of TEM-1 beta-lactamase with important secondary structure elements labelled. The omega loop is colored in cyan while catalytic residues are shown in magenta, including: Ser70, Lys73, Ser130, Asn132, Glu166, Lys234, Ala237. The green sphere shows the primary binding site for antibiotics. Figure Credit: [1].

molecule and S70 have been proposed. These involve a hydrogen bonding network between K73, E166, N174, K234. [81]

Although the omega loop has been well studied, the full extent to which it controls beta-lactamase activity has not yet been probed. In particular the role of dynamics in this flexible loop may yet explain the subtlties of substrate recognition.

The functional classification of beta-lactamase is based on which drug substrates that an enzyme can bind to. The three major classifications of beta-lactamase are as follows: 1) cephalosporinases, 2) serine beta-lactamases, and 3) metallobetalactamases. Each group is further subdivided to separate categories based on substrate recognition properties. For example, group 2 beta-lactamases are divided into penicillin resistant (2b), extended-spectrum resistant (2be), inhibitor resistant (2r), carbenicillin resistant (2c), cloxacillin and oxacillin resistant (2d), and carbapenem resistant (2f) subgroups. Additionally combinations of these subgroups can exist as well, e.g. example extended-spectrum cephalosporin and inhibitor resistant, 2ber.

Two of the most concerning functional classes of beta-lactamase are extended-spectrum beta-lactamase (ESBL) and Inhibitor Resistant beta-lactamase (IRBL). ESBL refers to beta-lactamase that exhibit resistance against third generation cephalosporins

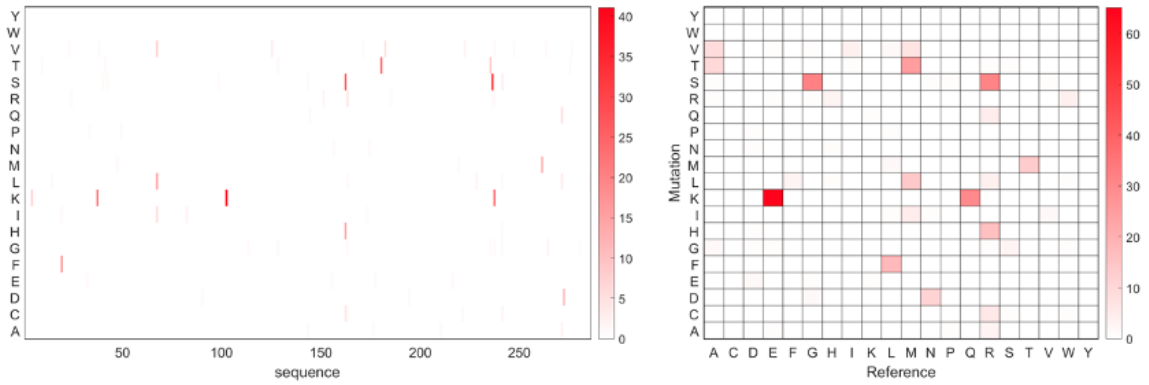

Figure 2.2: The mutational landscape of the 193 TEM beta-lactamase. The color scale is set by the frequency of mutation.

and other extended-spectrum beta lactam drugs, and ISBL refers to beta-lactamase that are able to inactivate antibiotics even in the presence of beta-lactamase inhibitors. ESBL and IRBL pose significant clinical concerns, with the CDC estimating that ESBLs alone were responsible for $197,000$ infections and $1.2 billion in healthcare cost in 2017, and these numbers are part of a rising trend. [15]

The resilience of beta-lactamase to new beta lactam substrates is a defining feature of these proteins and a key factor in their role as a primary cause of AMR. Various families of beta-lactamase have been evolved to confer resistance to specific classes of beta lactams, Examples include KPC (K. Pneumonia Carbapenemase) or OXA (Oxacillinase), whereas other enzymes in families like TEM or SHV can demonstrate a variety of different resistance phenotypes. [82, 83, 84] While structure is generally conserved among serine beta-lactamases, sequence conservation is specific family. This suggests that within families like TEM or SHV, the resistance profile is controlled by point mutations which modulate substrate recognition.

By aligning the sequences of TEM beta-lactamases, it becomes evident that mutations occur at relatively few spots along the sequence as depicted in Figure 2.2. This figure shows the mutation landscape for 193 variants in the TEM family. The distribution across the TEM sequence shows that mutations are not randomly distributed but occur in high frequencies at a small subset of specific loci. Additionally some mu-

tations, such as from E/K or G/S, have high frequencies across the family suggesting that these mutations provide the enzyme with the highest increase in evolutionary fitness.

Mutations can divided into a three main categories: primary, accessory, and unknown function mutations. Primary mutations directly change the substrate profile of beta-lactamase. Many of these mutations have been studied extensively in crystal structures of beta-lactamase, and are reviewed in [85]. In the TEM family, G238S and R164S are considered primary mutations conferring resistance to Cefotaxime and Ceftazidime respectively. Enzymes that confer resistance to third generation cephalosporins such as these are classified as ESBL, and these mutations are considered hallmarks of this phenotype. In Figure 2.2, the highest frequency mutations are often primary mutations.

Accessory mutations can affect catalytic specificity, however are usually act in conjunction with primary mutations to amplify changes. In TEM beta-lactamases, E104K and E240K are examples of accessory mutations. Specificity-changing mutations destabilize beta-lactamase, so additional accessory mutations like M182T have been found to restabilize the protein.

Finally there are mutations which do not change the specificity of the enzyme, nor have any obvious other impacts, but still are some how common within their respective families. One example is Q39K, the mutation that distinguishes TEM-2 from TEM-1. These two enzymes are known to have similar catalytic properties. Many TEM beta-lactamase can be considered as having been derived from a TEM-1 or TEM-2 lineage based on the presence of this mutation. A more recent study of this mutation in TEM-72 using molecular dynamics simulation [86] suggested that Q39K also acts to destabilize the protein.

2.2.2 Beta Lactamase Inhibitors

One of the reasons that beta lactam antibiotics remain effective treatment options, even with the emergence of beta-lactamase resistance, is the availability of beta-lactamase inhibitors. Traditionally enzyme inhibitors have been small molecules that are used in combination with a traditional beta lactam. These molecules come in two main generations. Currently most beta-lactamase inhibitors work by having their own beta lactam ring motif which binds in the beta-lactamase active site. Beta-lactamase inhibitors have been reviewed extensively in the literature [87, 81], and relevant information from these works are shown here.

The first generation of inhibitors includes clavulanic acid, sulbactam, and tazobactam. These inhibitors function by forming stable acyl-enzyme complexes with beta-lactamase blocking the enzyme from binding with a drug in the future. These reactions are irreversible and inhibitors that act this way are often called suicide inhibitors. These molecules have been clinically approved for use alongside particular antibiotics. For example, clavulanic acid is prescribed with amoxicillin or ticarcillin, sulbactam with ampicillin, and tazobactam with piperacillin or ceftolozane. Most penicillinase beta-lactamase are susceptible to these first generation beta lactamase inhibitors, and sulbactam even possesses its own mild antibiotic potency.

While beta-lactamase inhibitors have shown success in inhibiting penicillinases, the emergence of carbapenemase, cephalosporinase, and other beta-lacatamase resistance drives research into novel inhibitors. A second generation of inhibitors, which do not contain a beta lactam, has been recently developed. This includes avibactam (approved 2015 [88, 89]), vaborbactam (approved 2017 [90, 91]), and relebactam (approved 2021 [92]).

Avibactam works in combination with ceftazidime to restore efficacy against ESBL producing bacteria. Avibactam is a diazabicyclooctane (DBO) molecule noted for having a bridged bicyclic core. Although avibactam does not contain a beta-lactam

ring it still forms a stable Michaelis complex via a carbamate linkage with the catalytic serine in the active site of the protein. This complex, however, is far more stable, allowing the inhibitor to occlude ceftazidime from entering the binding pocket.

Vaborbactam and relebactam are inhibitors which can restore carbapenem efficacy against carbapenemase beta-lactamase. Vaborbactam has been approved for use with meropenem, and relebactam with a combination or imipenem and cilastatin. These inhibitor-drug combinations have reduced some of the threat from multi-drug resistant bacteria, in particular against carbapenemase beta-lactamases.

Vaborbactam is the result of extensive research into the utility of boronic acids as beta-lactamase inhibitors. They demonstrate efficacy against serine-like class A and C carbapenemases. Boronic acids have shown inhibitory properties against beta-lactamase, which involve interactions between the inhibitor and the catalytic serine of beta-lactamases. Relebactam is another DBO-based inhibitor with a bridged bicyclic urea core, derived from the structure of avibactam. Its mechanism of inhibition is similar to avibactam, however, it has been designed to be more stable, resulting in longer lived enzyme-inhibitor complexes.

An interesting case in the inhibition of beta-lactamase involves beta-lactamase Binding Protein (BLIP), a naturally occurring protein known to bind to beta-lactamases to inhibit their function. The BLIP/beta-lactamase protein-protein interaction (PPI) is a model system for studying PPIs because the it is well characterized and large. [93, 94] Due to this BLIP/beta-lactamase serves as a valuable model system for PPIs throughout literature.

The first BLIP was reported in 1996 [95] and since then three other BLIP or BLIP-like proteins have been discovered. [96, 97] Studies involving BLIP in complex with TEM-1 beta-lactamase show that the BLIP/beta-lactamase PPI surface could include more than 49 residues on TEM-1. [98] X-Ray Crystallography has been performed to directly resolve the structure of the beta-lactamase enzymes in complex with BLIP

[95, 99], and mutagenesis assays have been performed to determine the functional importance of various residues. [93, 100] It was found that residues 99-112, which form a helix-loop-helix region on TEM-1 that is adjacent to the binding pocket were most critical for BLIP binding. Interactions allow BLIP to cover the binding pocket of beta-lactamase, hindering drug molecules from accessing the catalytic Serine.

Although BLIP initially showed good inhibitory properties for beta-lactamase, it was also found to have a narrow spectrum of activity, primarily restricted to class A beta-lactamases. Additionally, as a larger protein, it may not have appropriate pharmacokinetic properties to be an effective inhibitor in a clinical setting.

2.2.3 Antimicrobial Peptides

BLIP stands out among beta-lactamase inhibitors because it is a protein rather than a small molecule like the other inhibitors discussed. Proteins and peptides represent a relatively new and untapped source for novel drugs. While synthetic antibiotic peptides are currently gaining traction, peptides have long played a significant role in the medical world. Insulin, first reported in 1922 [101], is a notable example of clinical peptides use. Since then only around 80 additional peptide-based medicines have made it to market. [102] Several factors have been attributed to the lack of success of peptide drugs, include: low oral bioavailability, short half-life, lack of target selectivity, and a lack of membrane permeability. [103] However, with advancements in synthetic peptide design, molecular engineers can now manipulate the pharmacokinetic properties of peptides with ease. These new capabilities addresses some of these prior issues, leading to a recent revival in the popularity of peptide drugs.

Most antimicrobial peptides (AMPs) are either naturally occurring peptides or derived analogs of natural occurring peptides. [104] Although few AMPs have become approved drugs, thousands of naturally occurring antimicrobial peptides have been discovered. AMPs are usually small peptides with around 10-100 residues, usually possessing a net positive charge. The positive charge arises from needing to counter-

act the negative charge present on bacterial membranes. Hydrophobicity is another important characteristic of these peptides. They often have partitioned hydrophobic and hydrophilic regions, with the hydrophobic portions of the peptide being able to penetrate into the bacterial membrane.

Peptide drugs may also be modified with the addition of non-natural amino acids and other small molecule motifs. This class of AMP is referred to as antimicrobial peptide conjugates, as the extra molecule is conjugated directly to the peptide. [105, 106, 107] The extra molecule may be another antibiotic and work in tandem with the peptide to increase its efficacy.

The work of Rudgers et al. in 2001 [108], which focuses potential peptide-based inhibitors for beta-lactamase, mined BLIP for potential peptide sequences which might bind with beta-lactamase. The study showed that just a fragment from BLIP was able to maintain binding efficacy with beta-lactamase. The fragments used were several short peptides selected from the crystallographically determined PPI surface. The selected peptides had the highest number of naturally occurring inter-molecular contacts. When synthesized and tested for binding affinity, the peptides exhibited significant binding, however the signal was weak compared to that of the full BLIP.

This work highlights the potential for designing of peptides for protein binding based on naturally occurring PPI surfaces. It should be noted that this process still required some by-hand modifications with the natural sequence, e.g. adding cysteine residues at the ends of the peptide to induce a disulfide bond to make the peptide cyclic. This was done with expert knowledge to improve the positioning of the peptide at the interaction surface.

2.3 Drug Discovery

Given these developments and the continued the prominence of beta-lactamase antibiotic resistance, there is still much progress to be made before the challenge is considered solved. Therefore, novel methods of drug discovery are needed to address

the current state of antibiotic resistance.

The current directions in drug discovery methods are reviewed in [109]. Traditionally, the focus was on naturally occurring substances such as with the discovery of penicillin. Modern drug discovery has evolved into a biochemistry-based pursuit for synthetic or natural molecules. This process can be divided into two parts: lead identification and lead optimization, where lead refers to candidate for a new drug. Lead identification involves the identification of leads, either by a screening process or by de-novo design. Traditionally, this requires many tedious experiments to determine which substances interact favorably with the target. [20]

After leads are identified, they are filtered for drug-like qualities to become drug candidates. A good lead is subjective of its target, however, Lipinski's rules, or the rule of five [110], can be used to ensure leads have general drug-like qualaities. Though the rules are not absolute, the core tennants are that viable drug-like substances will have fewer than 5 hydrogen bond donors, fewer than 10 hydrogen bond acceptors, a molecular mass less than 500 daltons, and an octanal-water partition coefficient less than 5. These criteria primarily address drug solubility and membrane permeability properties. The leads which pass these criteria can become potential drug candidates, where they leads are tested against the target, and optimized cyclically through an iterative optimization process. After a drug converges to its optimized state, it may then be passed along for the next phase to be approved for clinical trials.

High Throughput Screening (HTS) [20, 111] is often used during lead identification to search for small molecules that exhibit a high affinity for a target. This screening tests each molecule using highly parallel affinity experiments, however, this leads to an excessive time and cost for the development of each new drug approved for use. To alleviate these challenges, computational approaches which perform "Virtual Screening" have been integrated into this process.

Computational approaches, reviewed in [112], can be divided into two different

strategies: Ligand Based Drug Design (LBDD) and Structure Based Drug Design (SBDD). In LBDD, the target structure is unknown, but several ligands known to bind to the target are used to build and optimize a model ligand using Quantitative Structure-Activity Relationships (QSAR). [113] On the other hand, SBDD leverages the known target structure and exploits the properties of the target to search for drugs with favorable characteristics. This can be achieved by searching known small-molecule libraries, de-novo design, molecular dynamics/docking simulations, and, more recently, machine learning.

2.3.1 Peptide Drug Discovery

Peptides occupy a unique space in the pharmaceutical world, existing in the space between small-molecule drugs and large biologics. Naturally occurring peptides involved in cellular processes, including signal transduction, signal modulation, and innate antimicrobial defense, can be leveraged in the search for novel drugs. Naturally occurring and synthetic peptides used for medicinal purposes have a lower immunogenicity, and they can have high binding affinity and specificity for their targets compared with small-molecule drugs. Currently, the primary drawbacks of using peptide drugs include two key aspects: limited membrane permeability, impeding their access to specific targets, and a low stability *in vivo*. [114, 103]

The rise in interest in peptide therapeutics is in part driven by increased capacity for rational design and high-throughput screening of candidate peptides. Protein-peptide interactions exhibit remarkable specificity, and these interactions can be refined or manipulated through the introduction of point-mutations. In 1982 the first site-directed mutation studies demonstrated that mutations to a protein active site can lead to control over catalytic activities. [115, 116] Additionally, work in peptide-small molecule conjugates are expanding the possibilities of leveraging a wider array of chemical motifs to further control molecular interactions. Finally, with the advent of cost effective and more accurate sequencing methods, large arrays of peptides can

be tested at once.

Despite the advancements made in improving the viability of peptide drugs, de-novo drug design remains a time intensive and costly endeavor. In the absence of having a known binder for a potential drug target, a major challenge lies in identifying peptides that will interact well with the target. This challenge can be addressed through the use of computational methods.

2.3.2 Computational Methods in Peptide Drug Design

Computational methods have played a significant role in drug discovery, having a prominent role in modern SBDD and LBDD pipelines. These methods are described above.

To perform protein-based SBDD, structures of the proteins and peptides involved in drug interactions are needed. The three most common methods for determining high-quality atomic structures of proteins, as found in the PDB, include X-Ray crystallography, NMR, cryo-electron microscopy. [25] While these methods produce reliable structures of proteins based on experimental measurements, the need for structures in drug design outpaces the speed at which structures can be determined. Moreover, some proteins and peptides, such as IDPs, do not form long-term stable structures, posing additional challenges to the rational design process.

Various computational approaches to predicting structure exist, with varying levels of accuracy. Homology modelling [117], a popular approach, involves comparing a protein sequence with similar sequences of known structures and generating a protein structure using these templates. Other methods of computational structure prediction include fragment assembly and various physics-guided, de-novo structure prediction methods. [118]

Machine learning-based protein structure prediction has reached a level of acceptance within the biophysics research community that it can be used to complement structure determination by experimental methods. Recent advances in machine learn-

ing structure prediction are reviewed in [36]. Notably AlphaFold2 (AF2), released publicly by Google's DeepMind research group in 2021 [5], is at the forefront of this revolution in structure prediction, having decisively won the CASP-14 competition. [119] Since its release, AF2 has become an indispensable tool for protein design. While DeepMind did not return to CASP-15, AF2 continues to lead the field, with other labs improving on the model, making it more accurate and tractable to use.

Protein structure prediction models are not infallible, however having a plausible model for any protein or peptide to start a SBDD study reduces the barrier for conducting exploratory research. It is important to be aware of the limitations of structure prediction and, if possible, account for these shortcomings when designing any molecule. Despite the many successes of AF2, structure prediction and protein folding remain active areas of research and machine learning should not be used as a replacement for experimental verification.

Molecular docking is a computational approach which aims at predicting the how two molecules interact. Docking can be used for protein and small molecule or protein and protein interactions. Docking programs output predicted structures for a receptor/ligand complex, known as poses, and the relative strength of the interaction. The interaction strength is characterized by a scoring function which has some correlation to how well the two partners interact. A recent overview of docking methods and scoring functions is provided in [36].

From a biophysical perspective, the strength of interactions between two molecules is given by the change in free energy due to binding, ΔG_{bind}. This quantity encompasses many complex factors including intermolecular and solvation interaction energies. The value ΔG_{bind} can be indirectly calculated by molecular-mechanics Poisson-Boltzmann surface area (MM-PBSA), linear interaction energy (LIE), alchemical simulation methods, or umbrella sampling and Weighted Histogram Accumulation Method (WHAM), however these methods are unreliable, tedious to per-

form, and rely on slow molecular dynamics simulations to sample the docking process. [120, 121, 122]

To overcome this, modern docking methods have developed three types of scoring methods for computing an analog to free energy: physics-based scores, empirically-derived scores, and machine learning scores. [123] Although these scores correlate to docking affinity, they do not represent real physical free energy. Despite this, scoring functions with reasonable correlation to ΔG_{bind} can be used to rank docking pose candidates and filter large libraries to identify potential lead compounds.

Docking has been widely used to predict protein-small molecule interactions. However, the prediction of PPIs, and by association protein-peptide interactions, poses a greater challenge due to the inherent size and flexibility of proteins, which complicates to problem with numerous degrees of freedom. [124, 125]

Some PPI calculators incorporate the inherent flexibility of proteins to generate more realistic models. [126, 127] However, the gain in complexity due to this renders these methods intractable for high-throughput applications like virtual screening. Other docking programs perform rigid-rigid docking, where both the receptor and ligand are kept in a rigid conformation. [128, 129] Rigid docking is much faster than flexible docking at the cost of accuracy. This approach works well if both receptor and ligand are crystallized in their correct docking poses. Often this is not the case, and *in vivo* docking relies on the inherent flexibility of molecules to allow binding partners to take complimentary shapes.

Docking methods can provide terse rankings of PPIs for large scale studies, but should be validated by a more realistic docking procedure or experimental verification to ensure the accuracy of a predicted pose. Rigid-rigid docking in particular, is mostly suited for ranking candidates rather than analyzing biophysical interactions. Some programs, such as HADDOCK [127], use a hybrid approach. An initial rigid docking is performed to identify binding pockets on the target, followed by flexible refinement

using molecular dynamics to relax the molecules into a more favorable complex.

Despite these challenges, several key computational developments over the last decade have positioned peptide pharmaceutical design to emerge as a source for new medicines. GPU-accelerated computing has improved almost all aspects of computational structural biology. Many methods, especially docking and molecular dynamics, utilize grid-based algorithms that are highly suitable for GPU computation. For example, MEGADOCK is a software which uses GPU calculations and has cut down the time of protein-peptide docking prediction from minutes to seconds. Without GPUs, high-throughput docking studies would not be possible.

CHAPTER 3: METHODOLOGY/TECHNICAL ADVANCEMENTS

To investigate beta-lactamase function several methods were either developed or used. This chapter outlines the methods associated with this work and subsequent chapters present the results of these studies.

3.1 Molecular Dynamics

MD simulation is an invaluable tool for studying molecular interactions at the atomic level. Molecular processes across multiple timescales, ranging from picoseconds to milliseconds, can be investigated. [130] MD simulations use classical Newtonian physics to evolve a system over time. The physics simulated are defined by a force field, which parameterizes all interactions between elements in the system. [131, 132] The elements used in MD are usually atoms or groups of atoms in coarse grained models. The all-atom molecular mechanics forcefield uses quantum mechanical calculations to parameterize bonded and non-bonded interactions, including covalent bonds, electrostatics, van der Waals (VDW) forces, and the properties of atoms in their environment.

Simulations require a initial starting structure, which is modelled in real space, and usually obtained through crystallography or NMR. Prior to starting the simulation, the initial structure is spot checked for structural defects, such as missing atoms or residues, and prepared by solvating the protein or complex. Any additional ligands, cofactors or solutes are added and the whole system is neutralized with ions. The whole system is energy minimized using steepest descents so that the structure is in a physically realizable conformation. Energy minimization is crucial as any molecules added to the system, such as waters and ions, may not be initially placed

in a physically favorable position. Additionally structures from crystallography do not represent the true *in-vivo* native state conformation of the protein, and these defects can introduce error into the dynamics sampled in a simulation.

After minimization, the system is further equilibrated to a constant temperature and pressure (NPT) ensemble. First the system is coupled to a thermostat and equilibrated to a constant temperature. Second it is coupled to barostat and equilibrated to constant pressure. The equilibration step can be performed in several ways, however a common approach is to add position restraints to the protein backbone atoms to allow the waters and ions to relax and come into equilibrium with the protein, without significantly perturbing the protein's starting structure. Finally, the NPT ensemble is then freed for full production. The GROMACS MD software package is used throughout this work to simulate the motions of beta-lactamase. [133] GROMACS is a freely available and supports both MPI parallelization and GPU accelerations for optimized simulation performance. [134]

MD simulations were conducted to investigate the dynamics of beta-lactamase both in isolation and in interaction with various antibiotics. The wild type TEM-1 beta-lactamase, which is considered a diffusion limited enzyme, can reach catalytic efficiencies (k_{cat}/K_m) on the order tens of nanoseconds per reaction with the its appropriate substrate. [135, 136, 137] Accordingly, simulations lasting for a few hundred nanosecond should be sufficient to observe the dynamics associated with substrate recognition and binding. With a few interesting exceptions, beta-lactamase is well known to stably maintain its conformation. [79] This means that the majority of enzyme motion will be in side chain and backbone flexibility. Consequently, the choice of simulation time is thought to be long enough to observe most of the interesting protein dynamics of beta-lactamase in its native state. To increase the sampling of the systems of interest, a shotgun sampling approach was employed. Multiple simulations with independent starting structures were performed in order to account for

the randomness associated with system preparation and equilibration.

For this study, four mutants of beta-lactamase were selected: TEM-1 and TEM-2 representing WT resistance and TEM-10 and TEM-52 representing ESBL resistance. The TEM family was selected because it contains variants from multiple functional classes of beta-lactamase, including wild-type, ESBL, and IRBL conferring mutants. Additionally, the TEM family has both a highly conserved sequence and structure suggesting that functional differences in the proteins are likely to come from changes in protein dynamics.

Compared to the TEM-1 sequence, two primary mutations associated with ESBL are R164S and G238S. [85] The ESBL mutants used here were selected to represent each these mutation signatures. Specifically, TEM-10 [138] contains the R164S/E240K mutations while TEM-52 [139] contains the E104K/M182T/G238S mutations. TEM-2 [140] differs from TEM-1 by the Q39K substitution, This mutation is not directly responsible for increased substrate recognition, however computational studies have shown this mutation induces a synergistic destabilization of the enzyme. [86] The mutants used in the current study encompass a diverse representation of the "core" mutations within the TEM family that confer ESBL resistance.

To understand how protein dynamics facilitates function in beta-lactamase, particularly the effective binding to and inactivation of antibiotics, four drugs were selected to be modelled and simulated in complex with the different mutants. Two examples from both broad-spectrum and extended-spectrum antibiotics were selected.

Ampicillin (AIC) and Amoxicillin (AMX), derived from penicillin, are broad-spectrum antibiotics which are widely used across the world. They are susceptible to resistance by most beta-lactamase producing bacteria. The two drugs share a very similar structure. In AIC, a benzyl group is fused to the penicillin core, with an amino group attached to the linking methyl group. In AMX, the benzyl group is replaced with a phenol group. In recent years, drug/inhibitor combinations such as AMX and

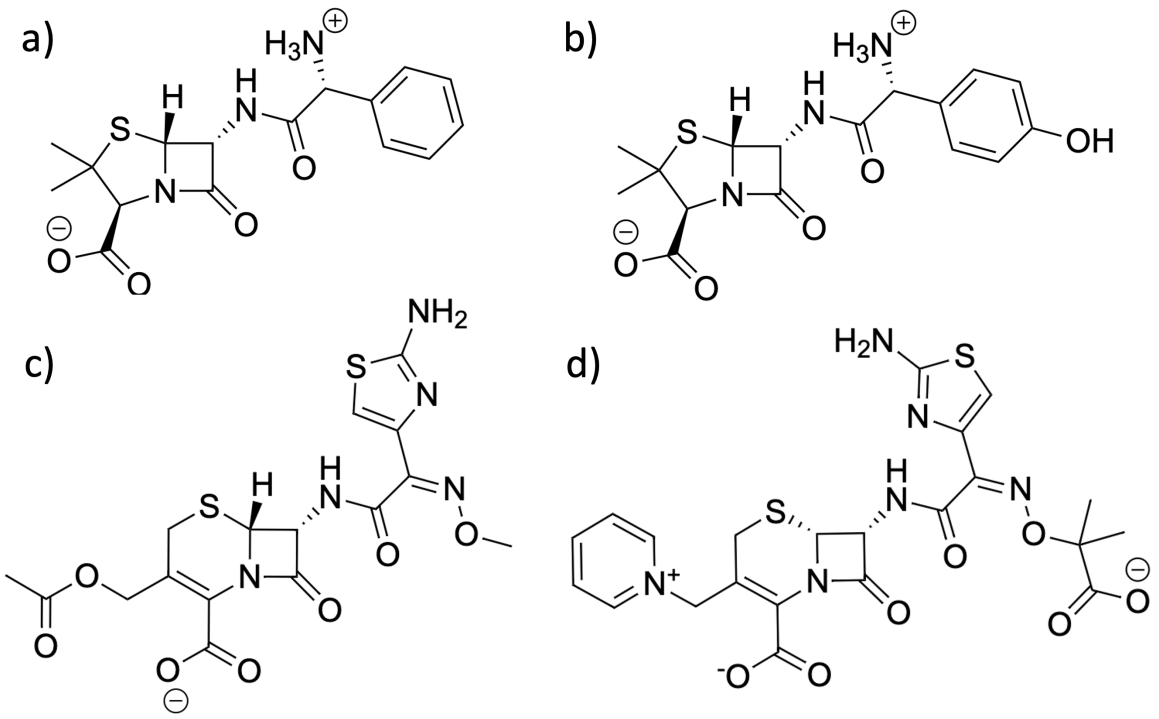

Figure 3.1: Structures of the four ligands used for simulation in this work: a) ampicillin (AIC), b) amoxicillin (AMX), c) cefotaxime (CEF), and d) ceftazidime (CAZ).

Clavulanic Acid have restored some of the effectiveness of these drugs.

Cefotaxime (CEF) and Ceftazidime (CAZ) are third generation extended-spectrum cephalosporin antibiotics selected for this work. The cephalosporin core structure allows for two functional groups. In CEF the first functional group attached to the core six-membered ring is a methoxyimino group with a thiazole ring attached and the second functional group is and O-COOH. CAZ has a similar thiazole plus methoxyimino functional group however the imino group has a more complex structure including two extra methyl groups and a carboxyl group. The second functional group is a charged pyridine ring. The bio-active forms of these molecules are shown in figure 3.1

3.1.1 Starting Structures and System Preparation

For the simulations conducted in this work, eight X-ray crystallography structures representing various TEM beta-lactamase were downloaded from the PDB. These crystal structures are summarized in table 3.1. To prepare the structures, crystal

waters and other artifacts such as co-factors or ligands were removed. When aligned by alpha carbon, the RMSD across all of the structures was 0.314 Å. This low value demonstrates how structurally rigid the TEM beta-lactamase family is, as each crystal structure had almost identical conformations.

In PyMol [141], the structures were computationally mutated to represent the TEM-1 sequence, and then further mutated and saved to express TEM-2, TEM-10, and TEM-52. In the case of crystal structure 1HTZ, the structure already represented a natural TEM-52 enzyme, and so for this case the experimentally derived structure was used instead of a computational mutation. There are not any examples of TEM-2 or TEM-10 in the PDB. The final dataset of protein structures contained 8 crystal structures mutated to 4 different enzymes for a total of 32 structures.

Table 3.1: Summary of beta-lactamase crystal structures used as starting structures for MD simulations.

PDB CODE	Resolution (Å)	Mutation (From TEM-1)	Reference
1ERM	1.7	N/A	Ness 2000 [142]
1ERO	1.7	N/A	Ness 2000 [142]
1ERQ	1.7	N/A	Ness 2000 [142]
1HTZ	2.4	E104K/M182T/G238S	Orencia 2001 [143]
1JWP	1.75	M182T	Wang 2002 [144]
1LHY	2.0	R244S	Wang 2002 [145]
1XPB	1.9	N/A	Fonze 1995 [146]
3JYI	2.7	N170G	Brown 2009 [147]

To simulate the selected beta-lactamase mutants in the presence of different antibiotics, a model of each ligand was constructed using base structures extracted from PDB entries. The entries used for each ligand are described in table 3.2. Initial models of each ligand were extracted their respective PDB files. These models were then processed in the molecular modelling software Avogadro [148], where bonds were reformed when necessary and explicit hydrogen atoms were added. Care was taken to make sure the ligands represented the correct charge state and protonation state for physiological pH to represent their the bioactive form. AIC and AMX both have

a neutral formal charge, however the carboxyl and amine groups are in their charged state. CEF and CAZ have a formal charge of -1, and CEF has a single charged carboxyl group while CAZ has two negatively charged carboxyl groups and a positive charge on the pyridine ring.

The process of deactivating beta lactam antibiotics by beta-lactamase can be approximated into three main steps: substrate recognition, acylation, and deacylation/turnover. [149] Each step involves the forming or breaking of distinct bonds, thus in each step the ligand has a distinct structure. This study is primarily interested in the dynamics which allow beta-lactamase to recognize different substrates, therefore the structures were designed to model the ligand prior to acylation.

Table 3.2: Summary of crystal structures from which initial ligand structures were found.

Ligand	PDB Code	Resolution (Ang.)	Reference
AIC	3KP3	3.2	Cheng et al. 2010 [150]
AXL	6I1E	1.64	Bellini et al. 2019 [151]
CEF	4PM5	1.26	Adamski et al. 2015 [152]
CAZ	5TWE	1.5	Patel et al. 2017 [153]

Protein ligand complexes were generated by docking the ligands into the apo beta-lactamase structures using global rigid receptor-flexible ligand docking in AutoDock Vina. [154] The combination of 4 ligands and 32 apo beta-lactamases resulted in a total of 128 docking experiments. AutoDock Vina ranked the poses by their predicted binding affinity, and top scoring poses were qualitatively examined to identify the poses which had the most favorable interactions between the protein and ligand. The specific critera used was that the ligand should be predicted to be near the known active site of the protein, and the reactive oxygen on the beta lactam ring was required to be facing the catalytic SER70 on beta-lactamase so that interaction would be more favorable. The most acceptable pose from each experiment was selected to be simulated.

All simulations were performed for 500 ns in an NPT ensemble at 300K and 1 barr

of pressure, using a base forcefield of Amber99SB-ILDN. [155] The parameters for the ligands were generated using Antechamber, part of the AmberTools package, and integrated into the forcefield with ACPYPE. [156, 157]

The simulation procedure began by placing the molecular models in a cubic box with minimum of 1.0 angstrom between the protein and the box edges. On average the simulation boxes ranged between 8 and 9 nm on each side. The box was filled with TIP3P water molecules and sodium ions were added to neutralize the net charge on the system. The whole system was energy minimized such that the maximum force on any atom in the simulation box was less than 1000 $kJ/(mol \cdot nm)$.

The minimized system was equilibrated with position restraints on the protein backbone. A velocity-rescaling thermostat was coupled to the system for 1 ns to equilibrate the temperature to 300 K. [158] Then a Parrinello-Rahman barostat was coupled with the system over an additional 1 ns to equilibrate pressure. [159] Finally the position restraints were dropped and production began.

To account for any artifacts of dropping the position restraints, the first 100 ns of each production run was excluded from subsequent analysis as a "burn-in" period. In total 160 independent MD simulations of beta-lactamase was run for this study, representing 20 distinct systems (combinations of TEM-1, TEM-2, TEM-10, and TEM-52 with ligation states APO, AIC, AXL, CEF, CAZ). Each system had 8 replicate 500 ns simulations.

Each simulation was inspected by downloading the raw simulation trajectory and visualizing it using PyMol. The simulations were monitored for notable events, such as large scale conformational changes or global unfolding events. None of the simulations showed major unfolding, although local unfolding of the terminal residues was observed, however this was not concerning as high flexibility on the protein termini are expected during MD. In several simulations, the omega loop was observed to partially dislodge itself from the main body of the protein and extend into the solvent.

These events, although rare, are of particular interest as they have been predicted to have potentially functional implications for beta-lactamase. Despite their rarity, the trajectories that exhibited these events were kept, as they represent physically realistic behaviors for the protein.

During simulations involving protein/antibiotic complexes, disassociation events between the protein and ligand were occasionally observed. This can occur for a few reasons, including due to a suboptimal starting pose or random fluctuations during the simulation which result in the ligand being knocked out of place. Since no external restraints or forces were used to force the ligand to stay bound to the protein, such events expected given the time scales of the simulations. However, in most of the cases, the inter-molecular interactions were strong enough to hold the antibiotic molecules on the surface of the protein. When dissociation did occur, the simulation was stopped and restarted from the beginning or, in the case of multiple failures, an alternate binding pose was selected.

Another event observed in some simulations was the ligand remaining in contact with the protein, and moving out of the binding pocket by "crawling" across the protein surface. When this happened, the ligand tended to get trapped on loops directly bordering the pocket or move across the beta sheet bordering the active site to occupy an alternate pocket between the the C-terminal alpha helix and the the loop between helix 10 and 11. This location appeared to be a common place on beta-lactamase where ligands seemed to have an affinity for residing. These events are further investigated in Chapter 4.

3.2 Essential Dynamics: JEDi

Protein motions are highly complex due to the large number of degrees of freedom associated with the x, y, and z positions of atoms. Even when coarse-grained to a single atom per residue, there can still be hundreds of degrees of freedom to explore. Several methods for understanding these motions were presented in Chapter

2, including ENMs and ED.

Java Essential Dynamics Inspector (JEDi) [54] is a Java-based software toolkit which provides a comprehensive set of features for performing ED analysis of protein dynamics. With JEDi, users can design a customized PCA analysis to gain insights into molecular systems, with a focus on generating easy to visualize and easily interpretable data. When a trajectory is loaded into JEDi as a set of PDB files, it calculates fundamental statistics up to the fourth statistical moment, and generates high quality plots for quick inspection. Additionally, RMSD and RMSF are computed and plotted for the trajectory. Most data calculated can either be returned as figures or in data files that can be exported for further analysis.

There are a variety of tools in JEDi which set it apart from other ED software. JEDi offers multiscale PCA in terms of coarse graining and atom/residue subsetting, allowing users to analyze protein motions at different levels. Coarse graining options include alpha carbon, backbone, heavy atom, and all atom, and subsetting allows users to specify subsets of atoms or residues for each coarse grained analysis. This multiscale approach to protein motion enables users to characterize the large global motions of proteins using a low-resolution level, while studying specific atomic level motions using a high-resolution level. This can include motions such as the movement of secondary structure elements or the arrangement of atoms within the active site of an enzyme.

A novel method of performing PCA on large sets of atoms called hierarchical PCA (HPCA) is also implemented in JEDi. In this method, PCA is first performed on all-atom representations of each residue independently. The the top N components from each residue are combined as low-dimensional global features of the protein, and PCA is performed again on this representation. HPCA relies on a divide-and-conquer approach to allow the all-atom ED of entire proteins to be calculated in a computationally efficient way. JEDi offers an array of advanced statistical options to increase

the robustness of an analysis, including covariance shrinking [160], sparsification, and outlier detection.

In addition to PCA specifications, JEDi offers several tools for high-quality analysis and motion comparison with a focus on easy visualization. In addition to high quality plots, PDB files are created, which can be visualized in PyMOL. For example, a PDB file is created where the B-factor column is replaced by the RMSF of the trajectory, which easily illustrates where large fluctuations are found. Additionally, JEDi will create a PyMol movies displaying the global motions described by each eigenvector and a movie displaying the superposition of all the essential motions. These high-quality movies allow direct interpretation of PCA vectors in 3d coordinate space.

JEDi also facilitates the comparison of essential motions between two proteins. The program provides a pooling tool which combines trajectories together for a single analysis. This is an important tool because the eigenvectors from independent PCA runs cant be directly compared unless the trajectories have all been aligned to the a common reference structure. JEDi performs such an alignment before running PCA.

By pooling the trajectories and conducting analysis, JEDi can identify and describe the differences in essential dynamics between two proteins. If the differences in motion between the proteins are larger than the largest motion observed in either protein individually, then JEDi will find these motions in the pooled trajectory. Alternatively, if multiple trajectories are aligned and JEDi analyzes each individually, then the eigenvectors can be directly compared in JEDi using the subspace analysis tool. This tool gives a quantitative comparison of essential dynamics by comparing the eigenvectors using overlaps and cumulative overlaps, and Root Mean Square Inner Products (RMSIP). [161, 162, 163]

Many other features have been included in JEDi, and a detailed discussion can be found in [54].

3.3 Allostery

A model for predicting dynamic allostery in proteins has been previously developed [41] and its application to the present work was described in [1]. This model uses perturbation theory to predict how the normal modes of a protein may change upon a simulated binding event. The normal modes of a protein can be computed from the Hessian matrix as described in Section 2.1.3.1. Using the quasi-harmonic approximation, the Hessian matrix, H_o, describing second derivatives of an effective potential V_{eff}, is approximated by the inverse of the covariance matrix, Σ, of the proteins motions. Using this, an unperturbed Hessian can be derived from a molecular dynamics simulation of a protein in its apo state as in Equation 3.1.

$$H_o = RT\Sigma^{-1} = kT \sum_{k=1}^{p} |k\rangle \frac{1}{\lambda_k} \langle k| \tag{3.1}$$

Here, λ_k is the eigenvalue associated with the kth eigenvector $|k\rangle$ of the covariance matrix, Σ. The number of eigenvector modes considered is p, which, for the exact inverse, would be equal to the full dimensionality of the protein or 3 times the number of atoms sampled. Using the quasi-harmonic approximation allows the Hessian matrix to be characterized without using an explicit functional form for the effective potential between atoms.

The modes with the lowest eigenvalues of the covariance matrix correspond to the smallest motions or random fluctuations of protein dynamics. However, because there are so many of these modes, they dominate the sum in Equation 3.1. The accumulation of so many small eigenvalues can introduce considerable noise when taking the inverse of the matrix, which originates from the instability of taking the inverse of very small values. To address this, an approach was developed in previous work to de-correlate this noise.

To de-correlate noise from the covariance matrix, first the eigenspectrum of the

matrix is obtained. Eigenvalues below a minimum threshold value, λ_m, are considered to be noise. To lessen the effect of this noise in Equation 3.1, the eigenvalues associated with the noise are replaced with the average value of noise eigenvalues, $\lambda_{avg} = \langle \lambda_i \rangle_{\lambda_i < \lambda_m}$. Using the average value preserves the trace of the matrix while removing the instability associated with taking the inverse of extremely small values. In prior work, the value of λ_m could be set as an algorithm parameter. However, here λ_m corresponds to the minimum uncertainty in atomic coordinates extracted from MD simulations, estimated to be 0.01 Å. After de-correlation, the Hessian in terms of the pseudo-inverse of the covariance matrix is given by Equation 3.2, where n_{max} is the number modes with eigenvalues above λ_m and p is the dimensionality of the system.

$$H_o = kT \left(\sum_{k=1}^{n_{max}} |k\rangle \frac{1}{\lambda_k} \langle k| + \sum_{j=n_{max}+1}^{p} |j\rangle \frac{1}{\lambda_{avg}} \langle j| \right) \tag{3.2}$$

The normal modes of the unbound protein can be found using the unperturbed Hessian matrix defined in Equation 3.2. To model the effects of effector ligand binding, a perturbation matrix, H_p, that models harmonic restraints to the local region around specific residues where the ligand is thought to bind, can be defined. These perturbations are modelled as additional springs having an associated spring constant k_s determining the strength of the perturbation. The exact form of the perturbation between two atoms i and j would be $V_{ij} = \frac{1}{2} k_s |r_i - r_j|^2$. When k_s is positive, the restraint represents a rigidifying perturbation to the protein.

In the allostery program described here, springs can be added to regions centered about specific carbon alpha atoms. Users can define the rules for adding springs, including the geometry, radius, and spring constant. To ensure the validity of using perturbation theory, the added spring constant should be small relative to the actual values of the Hessian matrix. The radius sets the size of the region in which springs can be added to, centered on the carbon alpha atom of the specified residue. The

geometry determines the pattern by which springs can connect atoms within this region, with a possible "triad", "ball", or "star" pattern.

The effective Hessian matrix, which combines the unperturbed Hessian with the perturbations, is given as $H = H_o + H_p$. Perturbation theory can be used to determine changes to the normal modes and their frequencies. To scan for allostery across the whole protein, the frequencies for each normal mode k can be computed using the unperturbed Hessian $\nu_0(k)$, with perturbations representing one ligand bound at the active site using one set of springs $\nu_1(k)$, and with perturbations representing an active site ligand and an effector molecule bound at a distal residue using another set of springs $\nu_2(k)$. The resulting $\Delta\Delta G$ can be calculated using the result of Cooper and Dryden, described in Equation 2.10. Practically, the lowest 6 modes should be dropped from this sum, as they represent the 6 trivial degrees of freedom for translational and rotational motions. The total change in binding free energy can be computed with Equation 3.3.

$$\Delta\Delta G = -kT \sum_{k>6} \ln\left(\frac{v_1(k)^2}{v_0(k)v_2(k)}\right). \tag{3.3}$$

The magnitude of $\Delta\Delta G$ found here is dependent on the strength of the perturbation. Therefore the values should not be considered as physical $\Delta\Delta G$ values, but rather a propensity scale that indicates potential allosteric signals. Furthermore, the models used in this work are built from coarse grained representations of proteins, which means that some significant vibrations in side chain motions could be missed. Finally, when scanning for allostery across an entire protein, if springs are added to the network such that springs at two binding sites are added in an overlapping way, then rigidifying contributions could be overcounted. In this work, the full beta-lactamase enzyme is scanned for allostery using Ser70 as the active site center. However no interesting signals were observed due to this overlapping effect, and in Chapter 4,

only the distal effects are considered.

3.4 SPLOC

Supervised Projective Learning with Orthogonal Completeness (SPLOC) [2] is a machine learning algorithm specifically developed for discriminant analysis of large multivariate time series data typical of molecular dynamics trajectories. A driving motivation during the development of SPLOC was that the nature of functional motions in proteins cannot be assumed, unlike in ED where the most significant motions are assumed to be the largest. SPLOC takes a supervised approach which assumes the dynamics that facilitate function in proteins will be shared by all proteins labelled as functional and absent in all proteins labelled as non-functional. SPLOC does not assume the nature of important atomic motions a priori,instead, it tries to maximize statistically significant similarities or differences between different classes of molecules presented. By focusing on these changes, SPLOC can elucidate the functional dynamics regardless of whether they correlate to large variance motions. SPLOC was designed to facilitate comparing the dynamics between protein mutants that have different functional properties through the analysis of MD simulations. The learning process in SPLOC employs a novel form of machine learning based on projection pursuit.

Input samples for SPLOC are partitioned into data packets, which represent ensembles of conformations of a protein. For training, the data packets can be labelled as functional or non-functional. SPLOC treats each data packet as one observation of protein dynamics. Rather than focusing on differences between individual protein structures from MD simulations, SPLOC focuses on maximizing differences between the emergent properties of trajectories. SPLOC optimizes a set of basis-vectors that discriminate data packets in feature space. In feature space, each data packet is represented by a $2N$ dimensional vector, where N is the number of SPLOC modes. The elements of this vector consist of the mean and variance of all the samples in the data

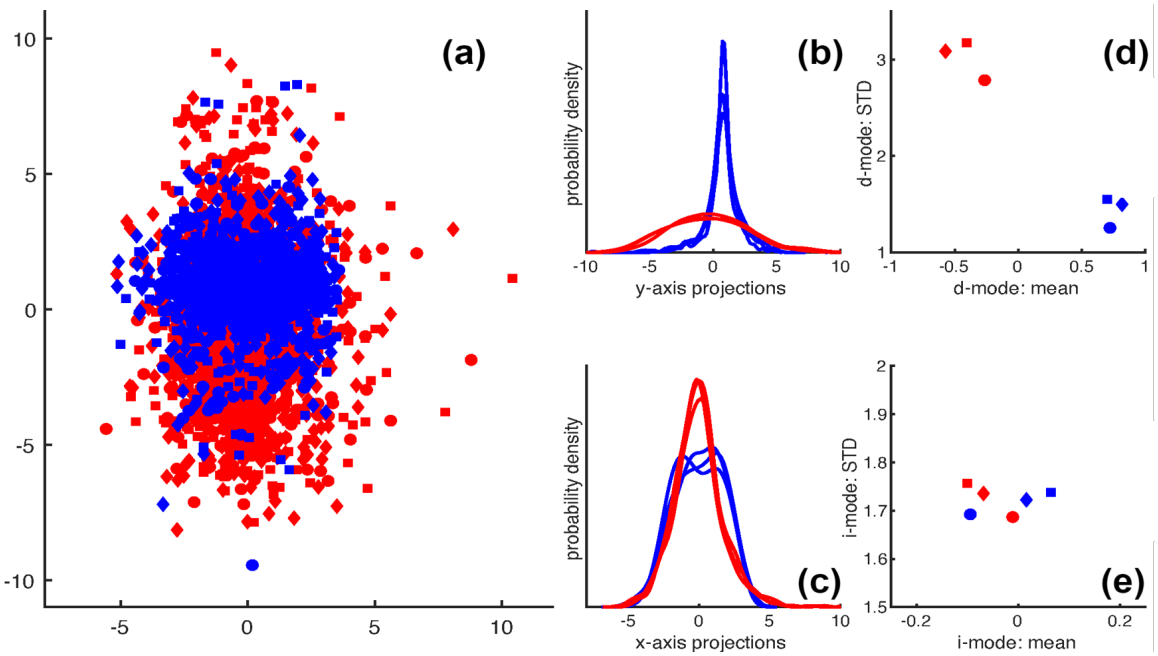

Figure 3.2: The connection between data packets and feature space in SPLOC is exemplified. In a), three data packets (squares, circles, diamonds) from two classes (red and blue) are shown projected into a 2D space defined by two basis vectors found by SPLOC. The x-axis represents an i-mode while the y-axis represents a d-mode. In b) and c), the distributions of the projections along both modes are shown. Finally in d) and e) the MFSP for both modes are shown, where the mean and variance of the distribtution of a data packet along the mode are plotted. Notably, the data packets differentiate by class along the d-mode (d) but not along the i-mode. Figure Credit: [2]

packet when projected onto each of the N SPLOC modes. Feature space is usually visualized in $2D$ slices per SPLOC mode, called a mode feature space plane (MFSP). A illustrative example of feature sapce is shown in Figure 3.2.

By considering data packets in feature space, statistical differences in data packets can be easily identified, even if the raw projections exhibit significant overlap. This emergent approach reflects the dynamic nature of proteins, where a single observation of a protein conformation is not sufficient to understand its dynamics. It is possible for two proteins with different functions to share 99% of their conformation space, while their functional differences are the result of the other 1%.

The optimization process in SPLOC involves maximizing a single score, called

efficacy, through a pareto-optimization of three scoring functions: selection power (SP), consensus power (CP), and clustering quality (CQ). SP measures how well the projections of data packets separate along a basis vector in respect to different classes via a signal to noise and a signal beyond noise ratio. This metric combines differences between the means and variances of projections. CQ is used to measure how well the data packets cluster in each MFSP. The CQ is used to weight a non-linear rectifier function that acts as an activation when computing the total efficacy of the mode. In this way SPLOC is able to learn the most appropriate orthogonal basis set for discrimination. Finally, Consensus Power (CP) measures the statistical significance for the discriminative power of the basis vectors across all functional-nonfunctional data packet comparisons.

The calculations for SP, CP, and CQ are performed individually for each SPLOC mode, and combined into a single value called mode efficacy. The total efficacy for a set of basis vectors is the sum of efficacy over all vectors in the set. This value is the objective function that SPLOC attempts to maximize. Depending on the value of SP and CQ, a vector is labelled as a discriminant (d-mode), indifferent (i-mode), or undetermined (u-mode) feature of the data. SPLOC considers both statistically discriminant and indifferent features as potentially interesting projections of the data, and as such both d-modes and i-modes contribute to mode efficacy. For this reason, SPLOC is incentivized to maximize the number of d- and i-modes that it finds.

For a basis vector to be identified as a discriminant or indifferent feature, it must meet particular criteria for each of the independent metrics. Otherwise, a basis vector will be labeled undetermined. To be a d-mode the vector must have high SP and CQ, while to be an i-mode, the vector must have low SP and CQ. Both d- and i-modes must have high CP to ensure statistical significance. The threshold for CP is called the voting threshold and can either be set by the user or calculated as a function of the sampling statistics of the training data packets. The exact implementation and

equations used to calculate SP, CQ, CP, and efficacy are described in the seminal publication, [2].

An initial guess for a basis vector set is either provided by the user, taken as PCA eigenvectors, or set to an identity matrix representing the original features of the data. The efficacy of the initial basis set is computed to begin the optimization process. Generalized Jacobi rotations within two-dimensional subspaces [164] are employed to rotate pairs of basis vectors until they maximize efficacy. Basis vectors are chosen for rotation by an importance sampling procedure that prioritizes vectors most likely to increase total efficacy of the basis. During training, SPLOC avoids getting stuck in local optima by employing undirected rotations in the undetermined space. This allows the u-modes to compete with already high efficacy modes in the discriminate or indiscriminate subspaces. The optimization continues until SPLOC converges to a solution.

SPLOC finds projection directions that represent different perspectives for understanding features which are either the same or different between two classes. This perspective can be biased with slight modifications to how SP, CP, and CQ are calculated. SPLOC offers several "modes" that allow it to bias solutions toward finding discriminant or indifferent features, or to allow an unbiased search for features. Using these modes SPLOC can be used to uncover different perspectives of the same data which highlight different information content. [165]

3.4.1 Functional Dynamics in SPLOC

In SPLOC the basis vectors, $|k\rangle$, form a complete set that is partitioned into subspaces as either d-modes, i-modes, or u-modes. Due to the completeness of the full set, each vector can be used to build a projection operator, P, that can isolate just the motions described by the particular SPLOC mode. Similarly, a filter for capturing the motions within each SPLOC subspace can be constructed using all the

modes in the space, as shown in Equation 3.4.

$$P = \sum_{k=1}^{N} |k\rangle \langle k| \tag{3.4}$$

When a data packet is projected into one of the subspaces by a projection operator, only the fluctuations described by the modes of that subspace are kept. This allows the discriminant and indifferent motions of a trajectory to be separated and analyzed independently. Subsequently, the RMSF can be computed in the original basis, giving the fluctuations in the discriminant subspace (dRMSF) and indifferent subspace (iRMSF) separately. This enables the identification of the exact locations on a protein where dynamic differences and similarities occur.

The SPLOC algorithm was originally designed to determine differences in molecular dynamics simulations of proteins. Remarkably, it has been shown to be an effective general discriminant analysis method for a variety of multivariate applications [2, 166]. SPLOC distinguishes itself from other dimensionality reduction and projection pursuit methods for several reasons. Firstly, SPLOC is a data driven supervised algorithm and does not rely on assumptions or underlying models to find similarities or differences between data classes. In contrast, many projection pursuit methods require a projection index serves as an underlying model for evaluating projection directions. In SPLOC, efficacy is used for this purpose, however efficacy is a data driven calculation. Secondly, by retaining all basis vectors as potentially interesting projections, SPLOC can predict both differences and similarities in the data. This, combined with the multiple bias modes, allows all of the data to be analyzed from multiple perspectives in order to fully understand the underlying properties.

3.4.2 SPLOC Discovery Likelihood

SPLOC also incorporates a method for computing the likelihood that data packets come from one class or the other. This soft classification method relies on probabilities

based on distributions along projection modes. This method involves presenting a known set of training data packets from classes 1 and 2 as $\{X_1\}_{N_1}$ and $\{X_2\}_{N_2}$, along with a set of d basis vectors for classification, $\{|k\rangle\}_d$. The goal is to compute the likelihood that a set of N_U unlabelled data packets, $\{X_U\}_{N_U}$, is likely to come from class 1 or not.

First, the probability distributions for each data packet, known and unknown, projected onto the kth classification vector, $f_Y(\langle X_Y|k\rangle)$ are computed. $\langle X_Y|k\rangle$ represents the projections of the data packet onto the basis vector, and Y can denote class 1, class 2, or unlabelled. The overlap integral between all pairwise probability densities, $I(X,Y)$, is calculated and mapped to a likelihood that the distributions are distinguishable, $P_d(I(X,Y))$.

The prior probability that all data packets between class 1 and class 2 are distinguishable along mode k can be computed as $w_1^{(k)} = \prod P_d(I(X_1,X_2))$, and the same can be done for class 2 to get $w_2^{(k)}$. The overlap integral between unknown data packets and known data packets from class 1 or 2 then represents the conditional probability that the unknown data packet belongs to that class when projected onto mode k, given as either $t_k(X_U|1)$ or $t_k(X_U|2)$.

Using Bayes theorem, the probability that an unknown data packet, projected along mode k, belongs to class 1 can be written as $p_k(X_U) = 1 - \prod(1 - t_k(X_U|1)w_1^{(k)})$. A value of 1 indicates the data packet is part of class 1. Conversely the probability that the data packet is not part of class 1 along mode k is written as $q_k(X_U) = 1 - \prod(1 - t_k(X_U|2)w_2^{(k)})$.

Finally, the classification is computed as the likelihood that the data comes from class 1 or not, $l_k(X_U) = p_k(X_U)q_k(X_U)$. Rather than computing if the data packet is class 1 or 2, this value is computed to reflect the concept of functional dynamics in proteins. Among all the motions a protein may undergo, only a specific motion may be required for function. Therefore any protein which has functional characteristics

must exhibit this motion. On the other hand, proteins that do not function have no restraint on what motions they can exhibit, except, that they cannot show the specific motion required for function. In other words, there is one unique feature that makes a class functional, but nearly infinite ways a system could not be functional. Considering this, it is logical to calculate likelihood to be class 1 rather than any other class. A total likelihood that the system comes from class 1, combining information from all modes, is called the discovery likelihood. It can be calculated by taking the root mean square of $l_k(X_U)$ over all modes in the classifying basis set.

3.5 Subspace Comparison

A common theme used between essential dynamics, SPLOC, and many other projection pursuit methods is the optimization of a complete orthonormal set of basis vectors, which can then be subdivided into subspaces describing interesting features of the data. When performing these analyses on different systems, various techniques have been developed to directly compare the information content in different vector subspaces. It is important to consider that in order to compare vector spaces they must have the same dimension, and the vectors must be defined in the same frame of reference.

Being in the same frame of reference means that the data should be observed consistently, ensuring that any two measurements are meaningful relative to each other. In the context of proteins, this requires aligning all the frames of each trajectory with each other in 3D space. This alignment procedure removes any relative global translational and rotational motion, establishing a consistent frame of reference for analysis.

When these two conditions are satisfied, there are several metrics for comparing subspaces, some of which have been included in both JEDi and SPLOC. Firstly, two individual basis vectors can be compared by computing their overlap. This is essentially just the inner product of two basis vectors, $|v_i\rangle$ and $|v_j\rangle$ as shown in

Equation 3.5, and directly measures the cosine of the angle between the vectors.

$$O(|v_i\rangle, |v_j\rangle) = \langle v_j|v_i\rangle \tag{3.5}$$

The overlap between identical, normalized vectors is 1, while perfectly orthogonal vectors have an overlap of 0.

To compare two vector subspaces, $V = \{|v_i\rangle\}_{N_v}$ and $U = \{|u_j\rangle\}_{N_u}$, the root mean square inner product (RMSIP) can be used to obtain a single number between 0 and 1 to describe how similar the vector spaces are. [163] A value of 0 indicates the subspaces are totally orthogonal, while a value of 1 indicates they are identical. The equation for RMSIP is valid for subspaces that are the same or different sizes, although they must be the same dimension. The equation for RMSIP is given in Equation 3.6.

$$RMSIP(U,V) = \sqrt{\frac{1}{max(N_v, N_u)} \sum_{|v_i\rangle \in V} \sum_{|u_j\rangle \in U} \langle v_j|v_i\rangle} \tag{3.6}$$

CHAPTER 4: FUNCTIONAL DYNAMICS IN BETA LACTAMASE

4.1 Background

The function of beta-lactamase is the hydrolysis of the beta lactam ring motif on antibiotics, thereby inactivating them. This fundamental function is shared across all classes of beta-lactamase. The complexity of this function increases when considering how various beta-lactamase are able to hydrolyze different beta lactam antibiotics at different rates, giving rise to the functional classes [73] such as penicillinases, cephalosporinases, and carbapenemases.

As more beta-lactamase have been discovered, it has become clear that some families, such as CTX and KPC, maintain their catalytic function across all members. However, other families, like TEM and SHV, express multiple functional classes. Interestingly, functional classes such as extended-spectrum or inhibitor resistance can arise in the same family multiple times through amino acid substitutions. Considering the case of the TEM family, these mutations often involve single, double, or triple point mutations, while overall, the sequence and structure remains highly conserved across the family. This suggests that control over substrate recognition is mediated by the biophysical effects of point mutations on the enzyme structure and dynamics.

This chapter focuses on the role of enzyme dynamics in controlling differences in catalytic efficiency in TEM beta-lactamase.

4.1.1 Research Goals

To explore the role of dynamics, a library of molecular dynamics trajectories was generated for several TEM beta-lactamase, representing either the wild type penicillinase/narrow spectrum cephalosporinase or extended-spectrum cephalosporinase

phenotypes. The details about the generation of these trajectories are described in Chapter 3.

This chapter analyzes the simulations using ED and SPLOC to identify the dynamics which exhibit the most notable shifts between different beta-lactamase systems. These shifts are either due to point mutations or substrate binding. Additionally, dynamic allostery is also considered using the in-house method for detecting dynamic allostery from Chapter 3.3 to identify potential allosteric hot-spots on the protein.

The goal of this research is to identify the pre-hydrolysis dynamics within beta-lactamase that are important facilitating protein-ligand binding. As the functional mechanism of beta lactam hydrolysis is conserved across all serine beta-lactamases, the observed differences in substrate recognition within the family must be due either to environmental effects or a physical mechanism that influences ligand capture and binding. Without major conformational change, protein dynamics provides a reasonable mechanism for doing this.

Protein dynamics encompass a wide range of motions, but often only a small portion of a protein native state conformational ensemble are relevant for function. The majority of motions in proteins are dominated by thermal fluctuations with little significance. For TEM beta-lactamase, which has a well conserved sequence and structure across its family, functional dynamics are likely to manifest as localized motions that optimize local conformational geometries or cooperative allosteric motions, rather than large global conformational changes.

This work investigates the changes in dynamics in beta-lactamase from main two perspectives: a protein perspective, focusing on the impact of mutations on beta-lactamase, and a ligand perspective, which aims to understand changes to beta-lactamase due to different ligand interactions. The protein perspective consists of comparing wild-type (TEM-1 and TEM-2) resistance vs extended-spectrum resistance (TEM-10 and TEM-52). The ligand perspective, meanwhile, compares the effects of

broad-spectrum antibiotics (AIC and AXL) vs extended-spectrum antibiotics (CEF and CAZ) binding to different beta-lactamases. In reality these factors may not be independent as shifts in dynamics or structure can be the result of the right mutations encountering a protein with the right antibiotic partner.

Finally, if discrimination between enzyme-ligand systems can be found on the basis of molecular dynamics, then it should be possible to predict whether a new type of beta lactamase will bind to an existing antibiotic by observing changes in protein motions. To test this idea, a classifier is built using SPLOC basis vectors that describe dynamic differences between proteins known to bind poorly or well with extended spectrum antibiotics. This application directly connects beta-lactamase function with enzyme dynamics, and illustrates a new application for using MD simulations to inform the drug design process.

4.1.2 Chapter Organization

The first part of this chapter presents a comprehensive analysis of the MD simulation library using commonly used analysis methods for molecular dynamics. Potential hot-spots of dynamic allostery, or regions on beta-lactamase which exhibit cooperative motions with the active site, in TEM beta-lactamase will be predicted using the apo simulations of beta-lactamase. As presented in [1], the initial results detailing how SPLOC can be used to identify functional dynamics in TEM beta-lactamase using MD simulation data will be shown. Next, this methodology will then be extended to understand beta-lactamase motion changes when interacting with different antibiotics. Finally, the concept of using protein dynamics to predict antibiotic resistance function in beta-lactamase will be prototyped and evaluated. In the final sections the broader context of these results will be discussed.

4.2 Molecular Dynamics of TEM Beta-Lactamase

A library of 160 simulations of beta-lactamase enzymes in various conditions were prepared for this work as described in Chapter 3. Each simulation was run for 500 nanoseconds, at 300 kelvin and 1 barr of pressure. To account for burn-in time after lifting position restraints, the initial 100 nanoseconds was discarded. In total, 20 unique systems were simulated, consisting of 4 different mutants either in their apo form and interacting with one of 4 different antibiotics. A shotgun-like approach was used to generate 8 replicate simulations per system, starting from 8 crystal structures. In total 3.2 microseconds of dynamics was produced per system.

For this work, JEDi was used to process the simulation data and conduct an initial essential dynamics analysis. A processing pipeline using JEDi was executed in three rounds to prepare the trajectories for analysis. Each part of the pipeline operated independently on each trajectory. The final processed trajectories can be pooled in various ways for further analysis.

To process each trajectory, first the frames of the associated GROMACS trajectory file is extracted to separate PDB files. Using JEDi, the preprocessing protocol extracts the coordinates from the PDB files and constructs an all-atom coordinate matrix. For this step, the starting structure of the trajectory was used as a reference structure for alignment by structural superposition.

In the second part of the pipeline, JEDi analysis was performed on the all-atom, heavy-atom, backbone atom, and carbon alpha atom coarse graining level, still using the initial structure as the reference. For each level JEDi produces a subset coordinate matrix. Since different mutants of beta-lactamase have different numbers of atoms, only the backbone and carbon alpha atoms are considered corresponding sets between all systems which can be aligned to a common reference. For all further analysis, the carbon alpha atom subset was considered.

In the final step of JEDi preprocessing, the carbon alpha subset of each trajectory

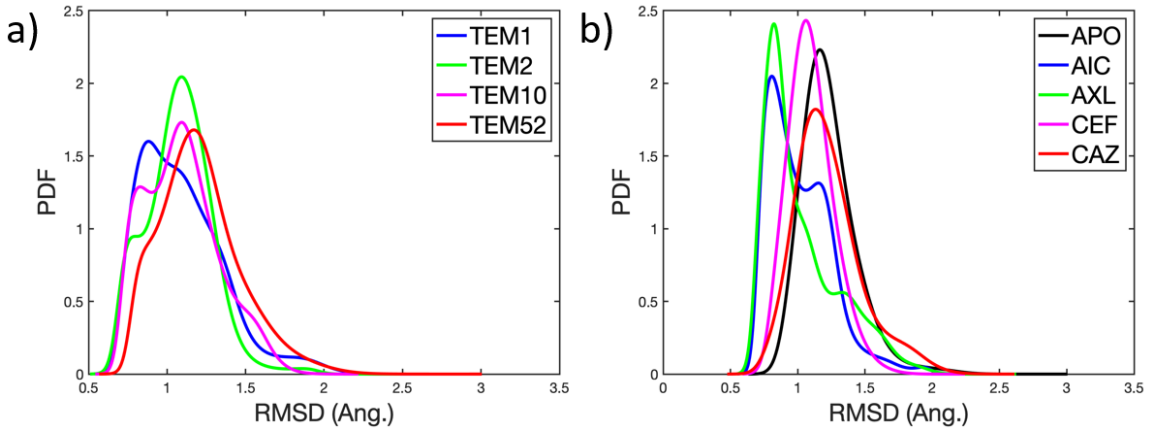

Figure 4.1: RMSD for molecular dynamics of beta-lactamase as pooled by a) which protein mutant is expressed or b) which ligand is present in the simulation. Apo is considered a separate ligand state.

was aligned to a single global alignment reference structure. Specifically, the first frame of the APO TEM-1 simulation starting from the 1ERM crystal structure, was selected. As a result of this pipeline, the 160 MD simulation trajectories were reduced to matrices containing the x, y, and z coordinates of only the carbon alpha atoms, all aligned to a common reference.

To assess how much change in the global structure of beta-lactamase occured during the simulations, the RMSD was computed for each aligned trajectory. The RMSD of for each simulation was pooled by either the protein mutant (TEM-1, TEM-2, TEM-10, or TEM-52) or the ligand state (apo, ampicillin (AIC), amoxicillin (AXL), cefotaxime (CEF), or ceftazidime (CAZ)) of the simulation. RMSD distributions over each group were found and plotted in Figure 4.1. The maximum RMSD across all simulations was less than about 3.0 angstroms, reflecting that in no instances did the beta-lactamase enzyme begin to unfold, and overall, the structure of the enzyme remained stable throughout all of the simulations.

RMSD values are commonly used to compare MD trajectories and assess major structural differences between proteins. In Figure 4.1, the RMSD values are centered at low values. When pooled by mutations, most simulations show significant overlap,

indicating that the simulations samples structures similar to their starting structure. When pooled by ligand, in Figure 4.1 b, the distributions show more pronounced differences compared to each other, however they still share a significant overlap.

Simulations involving either CEF, CAZ, or no ligand (APO) had slightly higher RMSD on average, which suggests that the proteins either had higher flexibility or were more malleable during their simulations. Additionally, this suggests that the proteins adopted a conformation during the simulations that was slightly different than the initial conformation.

The RMSF, obtained from JEDi, was also grouped in the same way as the RMSD and is shown in Figure 4.2 a-b. The RMSF curves for all systems are markedly similar, which indicates that on average regions of flexibility are conserved across TEM beta-lactamase enzymes. To more carefully compare the RMSF curves, Figure 4.2 c shows the difference in RMSF between TEM-2, TEM-10 and TEM-52 and TEM-1. Figure 4.2 d shows the difference in RMSF between AIC, AXL, CEF, and CAZ simulations compared to APO. The most notable changes in protein fluctuations occured at the omega loop (residues 163-178), particularly in TEM-52 enzymes and simulations including ceftazidime.

The small deviations observed in beta-lactamase crystal structures, along with the substantial overlap of RMSD and RMSF values in nearly all simulations, indicate that the enzyme did not undergo any significant global structural changes. This conclusion was verified by directly inspecting each simulation in PyMol.

To examine the global motions of beta-lactamase the essential dynamics of the proteins were analyzed. First, all the trajectories were combined in JEDi using the pooling functionality, and the essential motions were computed for the alpha carbon atom subset. To conserve on memory on the UNCC HPC, each trajectory was down-sampled by selecting only every 10 simulation frames. This downsampling resulted in a decrease of beta-lactamase snapshots from $1,280,000$ to a more reasonable $128,000$.

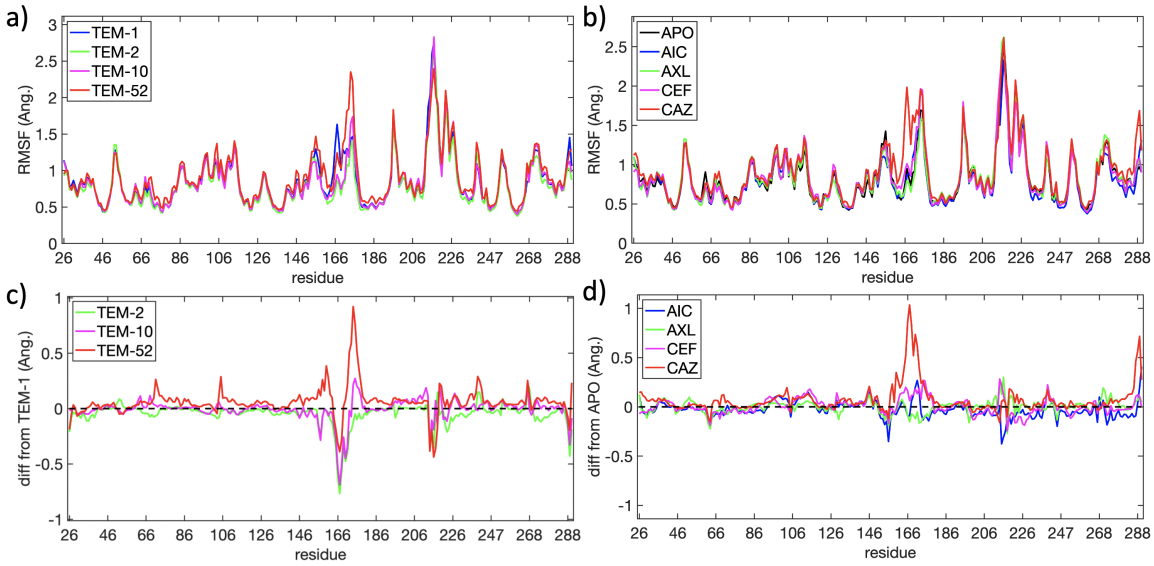

Figure 4.2: RMSF for molecular dynamics of beta-lactamase as pooled by a) which protein mutant is expressed or b) which ligand is present in the simulation. Apo is considered a separate ligand state. Differences in RMSF compared to TEM-1 are shown in c), and compared to apo in d).

No additional processing of the covariance matrix was performed. The results are shown in Figure 4.3.

Figure 4.3a illustrates the essential dynamics between all 160 simulations. All simulations generally overlap in a single continuous projection scatter. The marginal distribution for PC2 shows almost identical unimodal distributions per mutant, while PC1 displays a bimodal shape but is still highly overlapping. In Figure 4.3 b, the most variant motions across all simulations occur at the top of the C-terminal helix, and around residues 130-137 near the active site of the protein. Surprisingly, these are not the same regions that have the highest RMSF values occur as shown in Figure 4.2 a-b. These motions represent the most variant motions across all of the simulations, which may not always correlate to the largest motions in any particular simulation.

The essential dynamics of beta-lactamase are similar in all simulations. However, to quantify how similar the essential motions of each system were, JEDi analysis was performed on each of the 20 systems independently. Since all the trajectories have already been aligned to a common reference, the similarity between essential

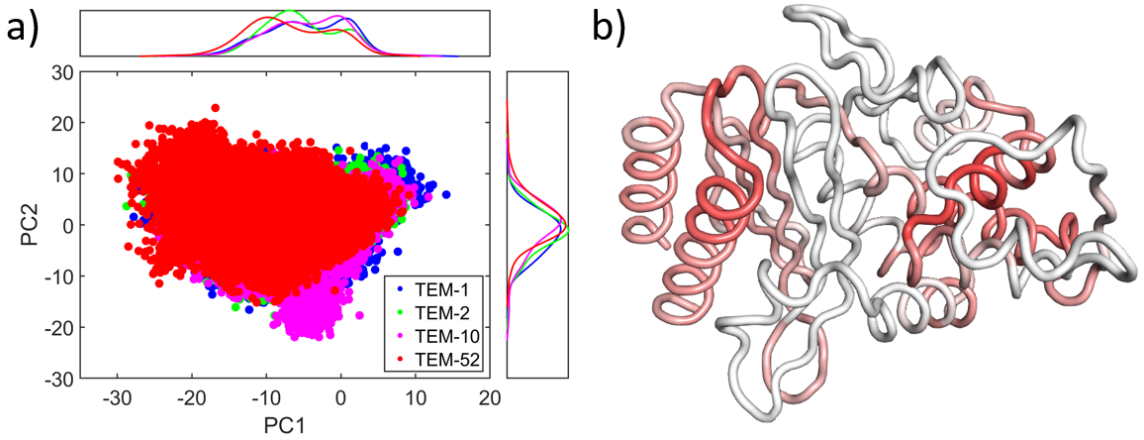

Figure 4.3: Essential dynamics of all beta-lactamase simulations pooled together. a) PCA projections for all simulation frames colored by what mutant was expressed in the system. The marginal distributions along each PC axis are shown along the sides of the plot. b) Beta-lactamase structure colored by where the largest global essential motions of the protein occur, as constructed by the top 10 PCA modes.

Table 4.1: RMSIP between 10/60 dimensional essential subspaces of protein dynamics compared by beta-lactamase mutant.

	TEM-1	TEM-2	TEM-10	TEM-52
TEM-1	1.00	0.821/0.921	0.817/0.919	0.801/0.904
TEM-2	0.821/0.921	1.00	0.840/0.918	0.797/0.899
TEM-10	0.817/0.919	0.840/0.918	1.00	0.826/0.908
TEM-52	0.801/0.904	0.797/0.899	0.826/0.908	1.00

Table 4.2: RMSIP between 10/60 dimensional essential subspaces of protein dynamics compared by antibiotic ligand present in the simulation.

	APO	AIC	AXL	CEF	CAZ
APO	1.00	0.828/0.902	0.817/0.904	0.831/0.917	0.826/0.913
AIC	0.828/0.902	1.00	0.837/0.914	0.827/0.911	0.764/0.900
AXL	0.817/0.904	0.837/0.914	1.00	0.835/0.920	0.799/0.901
CEF	0.831/0.917	0.827/0.911	0.835/0.920	1.00	0.826/0.917
CAZ	0.826/0.913	0.764/0.900	0.799/0.901	0826/0.917	1.00

dynamics subspaces can be quantified using RMSIP between subspaces spanned by the top N PCA eigenvectors.

Two methods were used to determine the number of PCA modes, N, to keep: Cattell's scree criterion [48] and cumulative variance. Cattell's scree criterion involves a qualitative examination of the scree plot, which plots the eigenvalues in descending order. The criterion for the optimal number of modes is where the the scree curve "elbows", or shows a significant change in slope. For the cases considered here, this was around 10 modes in each system. By cumulative variance, or the total percent of variance preserved when reconstructing the data using N PCA modes, it was found that at least 80% of the total variance in the systems was reconstructed by around 60 modes.

To assess the similarities of the essential dynamics, the RMSIP was calculated between all systems using 10 and 60 modes. Table 4.1 gives the RMSIP between enzyme mutant systems and Table 4.2 shows RMSIP for protein/ligand systems. In both cases, the RMSIP values are almost all very high, over .8, suggesting that the essential dynamics of beta-lactamase are very similar.

The similarity of global flexibility metrics and essential dynamics between either beta-lactamase mutants or antibiotic complexes supports the assumption that the physical mechanism of beta-lactamase catalytic function on different antibiotics is mediated through smaller localized dynamics. At the timescale of nanoseconds and microseconds probed here, beta-lactamase does not commonly undergo large conformational changes. In the context of beta-lactamase kinetics, this result is sensible, and it suggests that the fold of beta-lactamase enzymes has been optimized for recognizing the beta lactam motif and rapidly shuffling antibiotics into and out of the active site.

The functional differences in enzyme catalysis rates for different drugs does not follow a binary "resistant" or "non-resistant" scale, but rather it exists as a continuous

measure of how well a beta-lactamase can confer resistance to specific drugs. The functional differences lie in how easily different a beta-lactamase mutant can recognize and hydrolyze different classes of beta lactam drugs.

In the absence of conformational change, dynamics which mediate local conformational optimizations for different substrates are either induced by mutations in the protein perspective or induced by the presence of the substrate itself in the ligand perspective. The rest of this chapter presents more sophisticated approaches for uncovering these functional motions.

4.3 Dynamic Allostery in TEM Beta-Lactamase

In a previous study [1], all four beta-lactamase enzymes were scanned for dynamic allostery using the method described in Section 3.3. Initially, multiple values for the radius and spring constant were tested using all three geometries. As expected, the allosteric response curve demonstrated a roughly linear relationship with the perturbation strength. Through inspection, it was determined that the ball perturbation type, with a radius of 10 Å and a spring constant of 0.01 yielded the strongest signal. Using these parameters, the $\Delta\Delta G$ propensity curve was calculated for each of the 8 apo MD simulation trajectories per mutant. The average response across all 8 trajectories was computed, and error estimates were found using the standard error. The resulting responses are shown in Figure 4.4.

Notably, the raw propensity curves for each mutant were offset from the x-axis by roughly constant amount for each mutant. These offsets are likely due to incomplete sampling of beta-lactamase conformational space in the MD simulations. To make the comparisons between mutants easier, the offsets were subtracted from the curves in Figure 4.4a and are reported in the legend.

After subtracting off the offset, it is clear that the average signals for each mutant enzyme exhibits similar trends. This suggests that allostery is conserved across the TEM family, likely due to the conserved sequence and structure of the enzymes. The

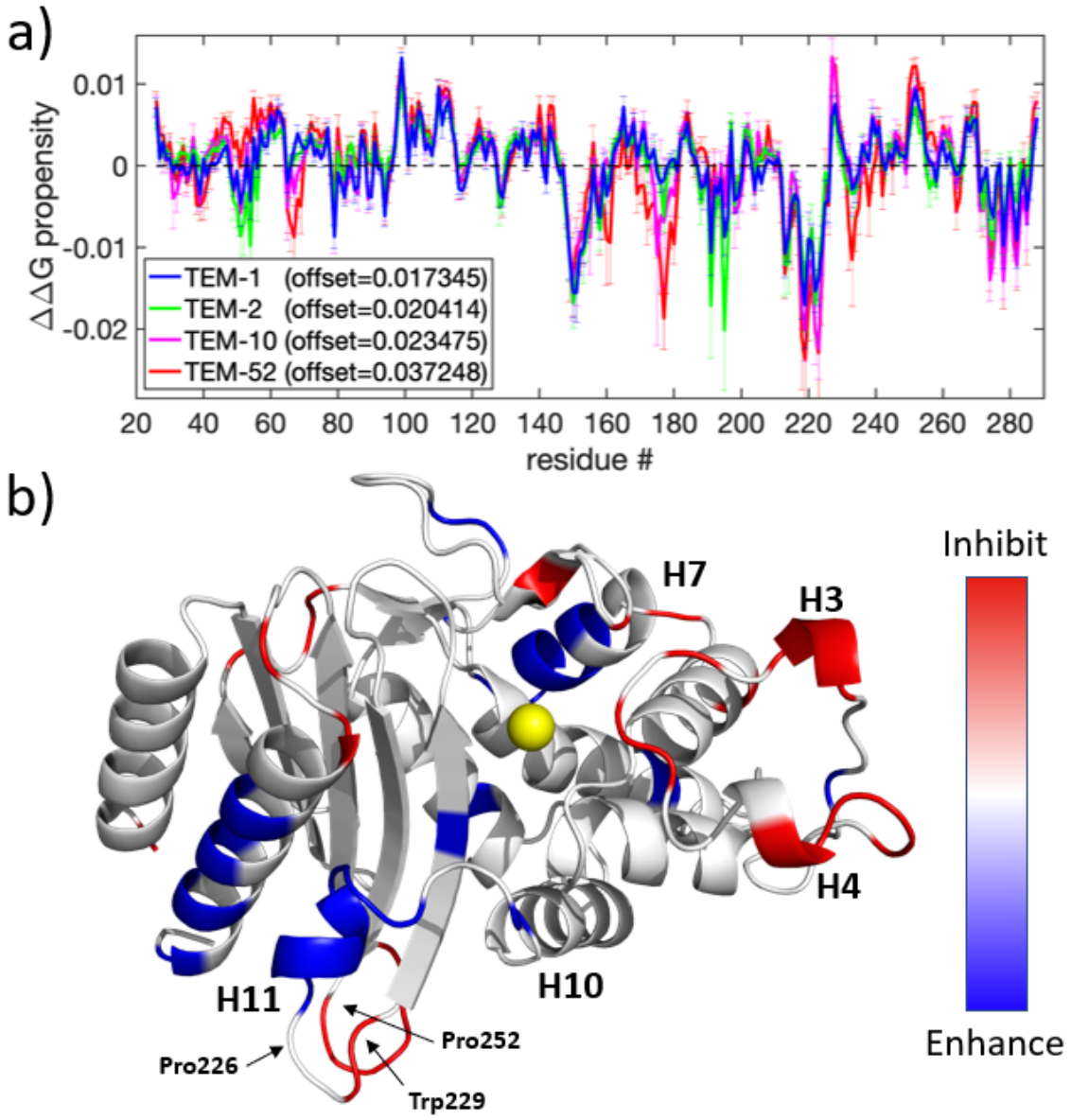

Figure 4.4: a) Allosteric response for TEM-1, TEM-2, TEM-10, and TEM-52 beta-lactamase across all residues. Error bars represent standard error over 8 independent simulations per mutant. b) Average allosteric response to rigidifying perturbations across all four mutants shows representative regions of high allosteric propensity. Signals under 0.0025 are zeroed and shown in white. Figure Credit: [1]

largest response is a negative allostery spike observed around helix 11. Although all four mutants display this spike, it is particularly prominent on TEM-10 and TEM-52. This result is interesting because it matches the location of a known allostery site discovered by Horn et al. [167] in 2004. Their research showed that this helix could undergo a confromational change, forming a pocket which allowed non-antibiotic molecules to bind to beta-lactamase. This binding subsequently decreased the binding affinity for antibiotics in the active site. The loss in affinity was attributed to a conformational change in Arg244, which is related to a destabilization in an aromatic ring stacking interaction by Pro226-Trp229-Pro252.

In the results of Figure 4.4 a, a stabilizing perturbation, which increases rigidity around this loop, was shown to increase binding affinity in the active site. This appears to agree with mechanism of the allosteric regulation described above, which suggests that destabilizing this region should decrease binding affinity in the active site.

Several other notable predictions of allostery in beta-lactamase were made in this scan. Due to the noisiness of the signal, this is best visualized directly on the structure in Figure 4.4 b. In this representation, the low value portion of the signal is removed, highlighting several regions of the structure with a strong allosteric propensity.

An inhibiting signal was found around helix 3 and 4 (residues 99-114), which are around 21 Å from the catalytic Ser70 (shown in the figure as a yellow sphere). Enhancing allosteric propensity can be observed around helix 7 (residues 150-155), about 18.5 Å from Ser70. Generally, these the signals appear to be conserved between all 4 mutants. However, TEM-52 does differ form the other mutants in some regions, particularly around residues 170-180 on the back side of the omega loop, where there is a large negative spike suggesting that a binding event here would enhance antibiotic binding affinity.

These findings are significant as they demonstrate that TEM beta-lactamase en-

zymes exhibit cooperativity, which is associated with binding efficiency at the active site. The allostery program was able to validate the known allosteric site on beta-lactamase, strengthening the likelihood that the other sites identified here represent novel allosteric targets for TEM beta-lactamase. The existence of an inhibiting signal at helices 3 and 4 are especially promising as it suggests that binding an effector in this region will decrease beta-lactamase ability to bind to antibiotics.

There is support in the literature that Tyr105, a residue on the connecting loop between the two helices, has a connection with beta-lactamase catalytic activity. [168] Mutation and NMR studies have shown that Try105 can take multiple conformations in TEM and CTX enzymes, one pointing towards and one away from the active site. [168, 169] Additionally, the NMR study correlated motion between Try105, catalytic residue Lys234, and Val216 between helix 10 and 11. A recent simulation-based study has also uncovered a dynamic connection between the motions of Ty105 and inhibitor binding at the allosteric site between helix 11 and 12 predicted by Horn [170] The results found here, in connection with the with literature, suggests that dynamic allostery may play a more significant role in beta-lactamase function than previously thought.

4.4 Identifying Functional Dynamics in TEM Beta-Lactamase

The research in this section is driven by the motivation that dynamics play a significant role in substrate recognition and catalytic function in beta-lactamase enzymes. Previously, in Section 4.2, differences between enzymes could not be described with global flexibility or essential dynamics. To further probe the protein dynamics, SPLOC is employed to compare dynamic changes between functionally different beta-lactamase under various conditions.

4.4.1 Identifying Functional Dynamics

The initial study presented in this section focused on demonstrating the utility of SPLOC for uncovering dynamic differences in proteins. These results have been previously published in [1].

At the time of this study, the full molecular dynamics library as described in 3.1 had not been completed, thus only a subset of the full library was used, consisting of all of the 32 apo beta-lactamase simulations (8 per mutant) and one 500 nanosecond simulation of each mutant/ligand complex combination. This gives a total of 16 holo simulations for a total of 48 simulations used in this analysis. For simplicity, this will be referred to the small set, S-library, while the larger library will be called L-library.

The efficacy metric in SPLOC, described in Chapter 3, takes into account both differences between data packets from different classes and similarities between data packets from the same class. In initial attempts to use SPLOC, it was found that using the trajectories from a particular class did not result in d-modes, as the data packets did not provide enough statistical similarity to satisfy signal to noise and statistical significance the algorithm.

To address this, a method was developed for bootstrapping data packets from a pool of trajectories from the simulation library sharing the same functional class. This was done to boost the observations-per-variable (OPV) in each data packet.

In this work, three methods for grouping simulations together were used: apo simulations, holo simulations divided according to the protein perspective, and holo simulations divided according to the ligand perspective. For a pool of N_{traj} simulations, each consisting of M frames of p variables, a subsample of m frames are randomly drawn from each trajectory and concatenated into a single data packet. Through this method, N_{DP} data packets are constructed with a total of $N_S = m*N_{traj}$ observations and an $OPV = N_s/p$.

If a class of trajectories is defined consisting of different subclasses of systems

Table 4.3: Description of data packets from different classes to be compared with SPLOC. Apo protein and holo protein classes in the protein perspective consist of TEM-1, TEM-2, TEM-10, or TEM-52, while holo protein classes in the ligand perspective consist of AIC, AXL, CEF, or CAZ.

	N_{traj}	M	m	N_{DP}	N_S	OPV
apo protein	8	8000	4000	16	32000	40.55
holo protein (protein perspective)	4	8000	4000	16	16000	20.25
holo protein (ligand perspective)	4	8000	4000	16	16000	20.25

grouped together (e.g. TEM-1 and TEM-2, both showing wild type resistance, or TEM-1+AIC and TEM-1+AXL with different broad-spectrum antibiotics), data packets in the class are constructed using observations solely from one subclass. This means that data packets are not biased in how they represent the dynamics of each subclass, while also allowing SPLOC to characterize fluctuations that may exist between subclasses. This approach was used unless unless otherwise stated.

4.4.2 Dynamic Changes Due To Mutations

First the effect of mutations on beta-lactamase dynamics was investigated by comparing each beta-lactamase mutant to the wild type TEM-1: TEM-1 vs TEM-2, TEM-1 vs TEM-10, and TEM-1 vs TEM-52. For each of these three comparisons, 16 bootstrapped data packets were constructed for each class being compared. SPLOC was run in its neutral mode of operation (MO) which does not bias toward either discriminant modes (d-modes) or indifferent modes (i-modes) and starting with an identity matrix as its initial guess. SPLOC was run a total of 10 times for each comparison to characterize the consistency among results, re-bootstrapping data packets each time. This procedure for running SPLOC was used in all subsequent sections, unless otherwise specified.

The average number of d-modes, i-modes, and u-modes found over the 10 replicate runs is presented in Table 4.4. SPLOC identified a consistent number of modes in each subspace as evidenced by the low standard deviations. The number of u-modes was small compared to d- and i-modes, indicating that SPLOC was able to partition

Table 4.4: Number of modes found for each apo protein comparison to TEM-1. Error is given as the standard deviation over 10 independent runs.

	# discr. modes	# undet. modes	# indiff. modes
TEM-1 vs TEM-2	93 ± 16.85	5.8 ± 3.32	690.2 ± 18.36
TEM-1 vs TEM-10	254.1 ± 22.5	48 ± 20.99	486.9 ± 26.54
TEM-1 vs TEM-52	239 ± 30.95	89.3 ± 16.26	460.7 ± 27.32

the motions easily. Notably, TEM-2, which is expected to be similar to TEM-1 on the basis of their similar catalytic properties, had the fewest d-modes and the most i-modes. This suggests that TEM-2 is more dynamically similar to TEM-1 than either TEM-10 or TEM-52.

Additionally, the MSIP (squared RMSIP) was computed between all subspaces (data not shown). In general, the discriminant subspace for TEM-10 and TEM-52 was conserved between replicate runs, however, for TEM-2, the MSIP for discriminant subspaces was on average lower. This means that SPLOC did not find the same d-modes each time it was run, supporting that differences between TEM-1 and TEM-2 are less pronounced.

Figure 4.5 illustrates the dynamic differences between beta-lactamase enzymes. As expected, the discriminant motions identified by SPLOC were not large compared to the total RMSF shown in Figure 4.2. TEM-2 had the smallest fluctuations in this subspace, as most of the motion was accumulated in the i-modes.

Comparatively, TEM-10 and TEM-52 had dRMSF values reaching between 0.6 and 0.8 Å. Both TEM-10 and TEM-52 exhibited large spikes at residues $194 - 201$ on the loop connecting helix 9 and helix 10 (H9-H10 loop). Smaller, but significant, spikes can be seen at the omega loop (residues $163 - 178$) and the primary catalytic residue Ser70.

The differences in dynamics observed in TEM-10 seemed to accumulate on the omega loop, whereas in TEM-52 the differences were distributed around the protein binding pocket. This observation partially reflects the spatial distribution of mu-

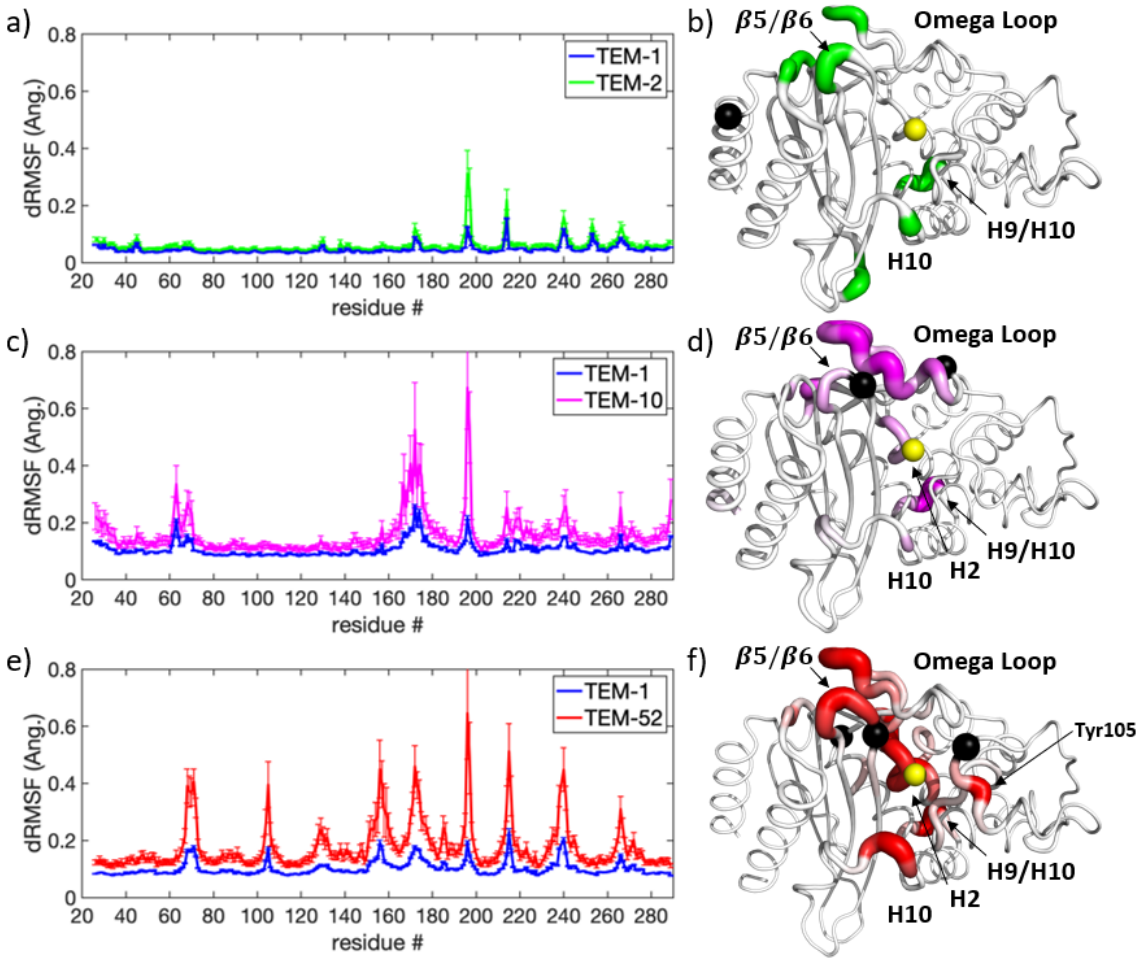

Figure 4.5: Dynamic differences between a-b) TEM-1 and TEM-2, c-d) TEM-1 and TEM-10, and e-f) TEM-1 and TEM-52. Panels a,c,e) show the dRMSF and panels b,d,f) show the dRMSF projected onto a structure of beta-lactamase. The unique set of mutations for the non-TEM-1 enzymes are shown in black and the catalytic Ser70 is shown in yellow for each. Figure Credit: [1]

tations in these two beta-lactamase mutants. In TEM-10 both R164S and E240K are on or proximal to the omega loop, while the mutations on TEM-52, E104K and G238S, are on opposing sides of the binding pocket. The final TEM-52 mutation, M182T, is on the complete backside of the protein. It appears from this observation, for TEM-10 to accommodate its preferred extended-spectrum cephalosporin ligands the mutations change the motion of the omega loop, while the mutations on TEM-52 induce motions all around the binding pocket, including on the alpha helix holding Ser70.

These results align with several speculations that have been proposed in the literature regarding how ESBL enzymes increase their substrate specificity for extended-spectrum cephalosporins. In general, the omega loop has been identified as a key component of substrate selectivity. [171] The results here show that different mutations can induce different changes in motions in beta-lactamase.

For the G238S mutation found in TEM-52, it has been suggested that the mutation causes the omega loop to be able to push away from the β5-β6 turn either from steric interactions or the loss of a contact with Asn170 [172, 173]. In this work, increased flexibility is found in TEM-52 in both of these structural regions supporting this mechanism. Notably, in the expanded L-library, where the omega loop was observed, albeit rarely, to push itself away from the rest of the protein. The the rest of enzyme was able to maintain stability during this event, and in some cases the omega loop was able to return to its original conformation. If this mechanism plays an important role in modifying substrate specificity, the motion will likely be more common on timescales beyond what was probed in this work. Increased flexibility was also observed in other loops surrounding the active site in TEM-52, which suggests that the widening of the active site is not exclusive to just the omega loop and β5-β6 turn.

The mechanism of R164S has been associated with a change in conformation of

the omega loop in several studies. [174, 137] It has been suggested, based on crystal structure observations, that the loop can open up and form a cavity which allows the larger cephalosporins to access the binding pocket. [144] In Figure 4.5 c-d, the changes to the R164S containing TEM-10 are strongest at the omega loop, which supports this mechanism of action. Glu166 on the omega loop is noted to be involved in the acylation/deacylation process of beta-lactamase. In TEM-2 and TEM-52 the changes to the omega loop were constrained to the turn not containing this residue, suggesting that it is important for these enzymes to maintain the proximity of Glu166 to the active site. On the other hand, in TEM-10 this dynamic change is across the whole front side of the loop. This might suggest that the acylation/deacylation may favor Lys73 as a general base instead of Glu166 in TEM-10. [175, 176]

One of the most interesting results shown here is the large spike at the H9-H10 loop. This loop is both very distal to the binding site on the protein and occured consistently in each comparison, including TEM-2. The literature regarding this region of the protein is sparse, the loop's connection with the function of beta-lactamase is unknown. Prior studies have noted that catalytic properties of beta-lactamase are resilient to mutations here, with the exception of Leu199. A later study showed that the L201P mutation had a minor stabilizing effect on the protein similar to M182T on TEM-52. [177, 178] A recent study on allosteric pathways suggested that this loop acts as a focal point for passing allosteric signals across the enzyme. [179]

4.4.3 Dynamic Changes Due To Ligand Binding

Next, the trajectories in the S-library were divided by enzyme mutant, and each group was further separated into classes of apo or holo. This comparison aims to identify where dynamics shift upon a ligand binding to the protein. Data packets were constructed using the bootstrapping method, however, due to the limited holo trajectory data available, frames from simulations of different subclasses (TEM-X + AIC, AXL, CEF, or CAZ) were mixed when constructing holo data packets. Just

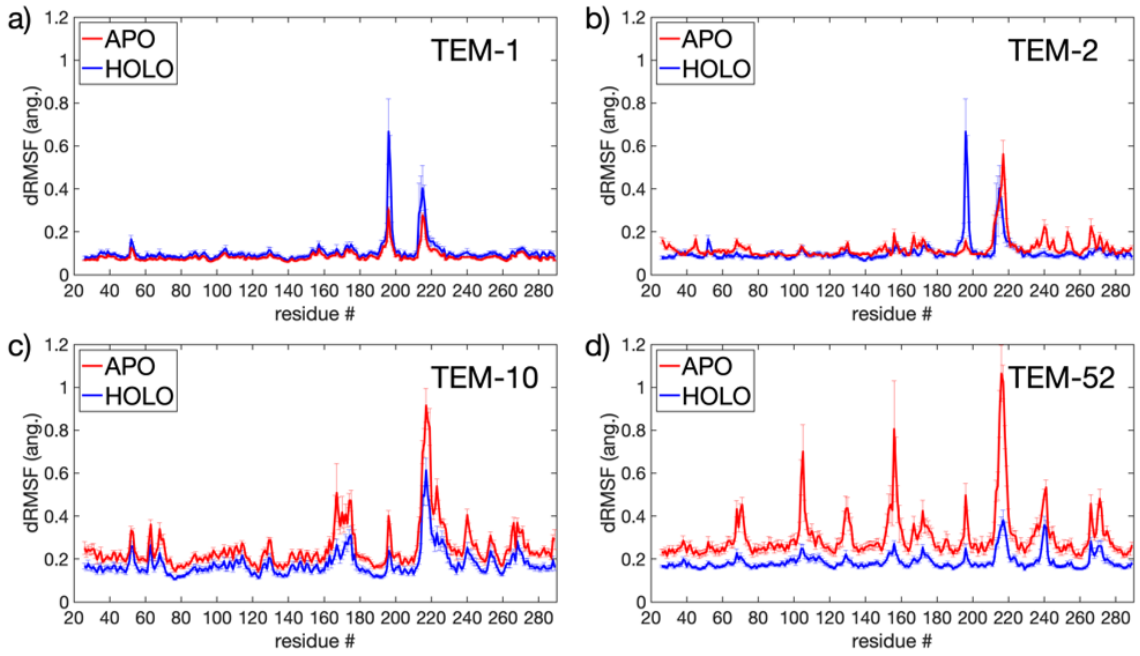

Figure 4.6: Average dRMSF for comparing apo vs holo simulations for a) TEM-1, b) TEM-2, c) TEM-10, and d) TEM-52. Figure Credit: [1]

as before, SPLOC is run for 10 times per comparison in the unbiased MO, using an identity matrix as its initial guess.

The average dRMSF for these four comparisons, apo versus holo in TEM-1, TEM-2, TEM-10, and TEM-52 enzymes, are shown in Figure 4.6. In the case of the wild-type resistant TEM-1 and TEM-2 beta-lactamase (Figure 4.6 a-b), the dRMSF response was small, with spikes only occurring in two locations: the H9-H10 loop and around H10-H11 loop. For extended-spectrum beta-lactamases TEM-10 and TEM-52 (Figure 4.6 c-d), the response is more distributed around the protein.

TEM-10 shows that ligand binding appears to reduce motion in the omega loop, characterized by the dRMSF in the apo enzyme being higher than the holo, and induce a shift in motion around the H10-H11 loop, shown by the tall dRMSF spikes for both apo and holo. In TEM-52, ligand binding appears to reduce motion at multiple sites across protein. Most of these sites appear centered on the binding pocket, including the catalytic Ser70, the H3-H4 loop, the end of helix 7, and the

H10-H11 loop. Notably, the omega loop does show more flexibility in the apo enzymes for both TEM-10 and TEM-52, however this peak is much less significant compared to the other peaks.

The location of the dynamic changes on TEM-52 during ligand binding align with the dynamic changes in the apo protein resulting from mutations. This suggests that the increased flexibility induced by mutations in the binding pocket is directly related to regions where antibiotics are able to anchor themselves to the protein. Conversely, the largest shift in motion for TEM-10 occured at the H10-H11 loop. However, the dynamic shifts resulting from TEM-10 specific mutations did not impact this region as much. This suggests that the change in motion in TEM-10 is dependent on ligand binding. More generally, ligand binding in different beta-lactamase enzymes may be occurring by different mechanisms.

Spikes in discriminant motion occurred at both the H9-H10 loop and around the H10-H11 loop in all four mutants. The role of flexibility in the H10-H11 loop is more evident, as this loop borders the binding pocket of the molecule and potentially directly interacts with the ligand. This flexibility could be harnessed by the protein to help stabilize ligands of different sizes and shapes during hydrolysis. Additionally, the observation that this loop appears to change its motion due to ligand binding, rather than due to mutations, supports this idea.

The H9-H10 loop shifts its dynamics due to both mutations and ligand binding. This loop forms a hinge on the backside of the protein, directly linking the omega loop and helix 10, both regions that border the active site and also show significant changes in motion through SPLOC. Consequently, the introduction of flexibility in this loop may serve as a general mechanism for beta-lactamase enzymes to relieve structural tension arising from enzyme mutations or environmental factors. This would help beta-lactamase be more resilient when encountering new antibiotics.

Finally, an attempt was made to understand the common dynamic characteristics

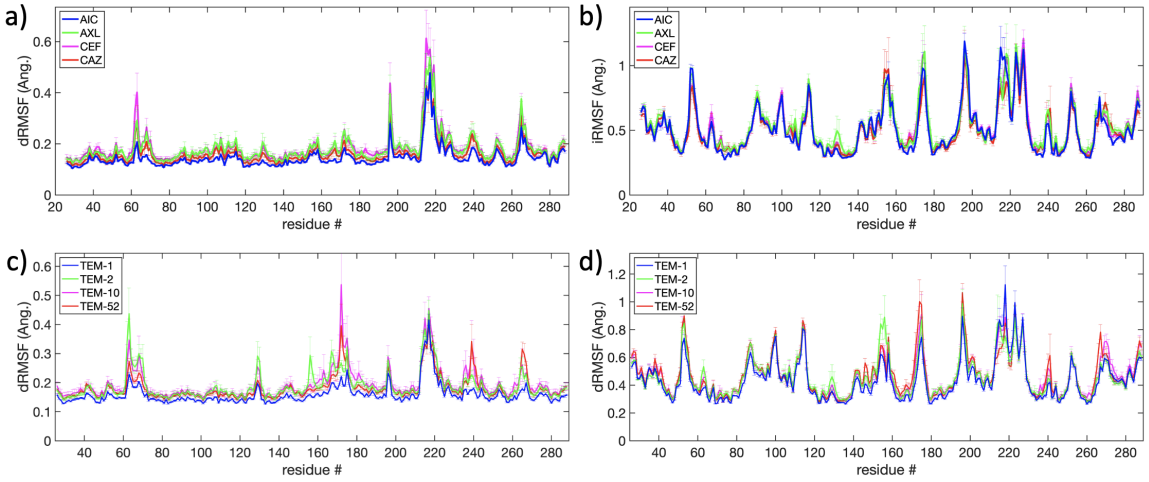

Figure 4.7: dRSMF and iRMSF from comparing holo beta-lactamase trajectories in the a-b) ligand perspective and c-d) protein perspective. Figure Credit: [1]

of beta-lactamase bound to different antibiotics. The prior work comparing apo and holo simulations showed that ligand binding induces either a quenching or shift in motion at specific locations on the enzyme to accommodate the antibiotic interactions. To compare motions among beta-lactamase enzymes bound to different ligands, the 16 holo trajectories in the S-library were split into two classes based on either the protein or ligand perspective.

In the protein perspective, simulations containing wild type-like enzymes (TEM-1/TEM-2) and those containing extended-spectrum enzymes (TEM-10/TEM-52) were grouped together. In the ligand perspective, simulations containing broad-spectrum antibiotics (AIC/AXL) and containing extended-spectrum cephalosporins (CEF/CAZ) were grouped together. Bootstrapping was employed as shown in Table 4.3 to construct data packets. SPLOC was employed to compare the two classes in the same method as before, running 10 trials in the unbiased MO, using an identity matrix as the initial guess.

The dRMSF and iRMSF from comparing the holo simulations in both the ligand and protein perspectives are shown in Figure 4.7. In the ligand perspective, dRMSF (Figure 4.7a) shows that changes in motions due to different classes of antibiotics were

observed at the H10-H11 loop. This loop exhibits discriminant fluctuations in all four subclasses from the apo vs holo analysis, which suggests that this loop is always fairly flexible, however, exhibits different motions depending on the ligand present. This supports the previous SPLOC results described above, that this loop does interact with antibiotic ligands as they bind, however the nature of this interaction is ligand-specific. In the protein perspective, several regions, including near the catalytic Ser70, the omega loop, and the H10-H11 loop, display different motions. The omega loop in particular displays higher fluctuations in ESBLs.

It is worth noting that both SPLOC analyses shown in Figure 4.7 utilized the same pool of trajectories, but partitioned in two different ways. Consequently, different discriminant motions were identified. The differences between the protein and ligand perspective dRMSFs highlight how examining the same simulations from different perspectives can change what motions are considered important. Training SPLOC using all available data in this manner lets SPLOC generate hypotheses to explain any differences between the groupings that are observed. This highlights the exploratory ability of SPLOC to identify functional dynamics, while other methods like ED may struggle. Considering the size of the fluctuations in the indifferent subspaces, PCA would have likely not been able to identify the discriminant motions against the background of larger fluctuations. This is a major advantage of using supervised learning methods like SPLOC for comparing MD simulations. [165]

4.5 Functional Dynamics of Substrate Interaction in Beta-Lactamase Using SPLOC

In this section, the L-library of MD simulations, which includes equal sampling across all 20 beta-lactamase systems (four enzyme mutants and five ligand states including apo), is used to assess changes in dynamics due caused by ligands. For the following work, the bootstrapping approach described in Section 4.4.1 was adjusted so that each bootstrapped sample contained the same number of total observations as

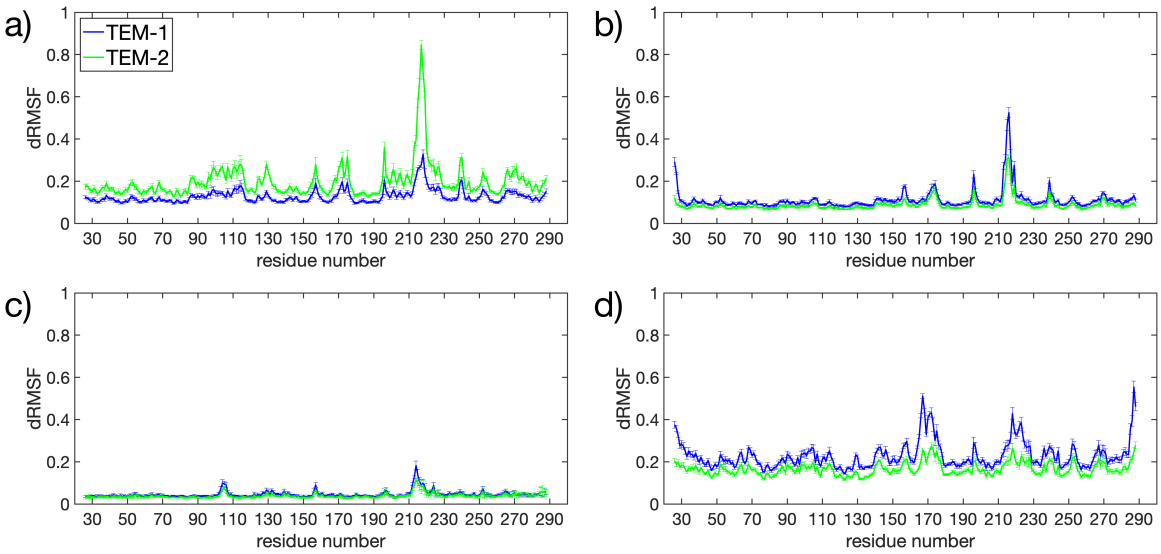

Figure 4.8: Changes in dynamics between TEM-1 and TEM-2 while in complex with a) AIC, b) AXL, c) CEF, d) CAZ ligands.

each raw MD trajectory. This adjustment was made to demonstrate that discrimination between proteins at OPVs similar to unprocessed simulations is feasible. Using the L-library, N_{traj} is now 8 for holo simulation classes, m is now 1000, and N_{DP} is 15. This results in $N_s = 8000$ and an $OPV = 10.14$ per data packet. To compare the new bootstrapping method with the previous one, the apo comparisons were redone using the new data packet construction method.

Qualitatively, the dRMSF from both methods appeared to be similar. To quantitatively compare these results, the RMSIP was compared between discriminant subspaces obtained from both methods. On average, the RMSIP between discriminant subspaces for the (old/new) method was (0.653/0.614), and the average RMSIP between the discriminant subspaces from different methods was 0.57. This result suggests that, on average, any pair of discriminant subspaces found by different bootstrapping methods are as similar to each other as another pair of discriminant subspaces found using the same bootstrapping method. Therefore, it is concluded that the new bootstrapping method does not compromise the quality of the SPLOC results, but with the advantage of using many less observations per data packet.

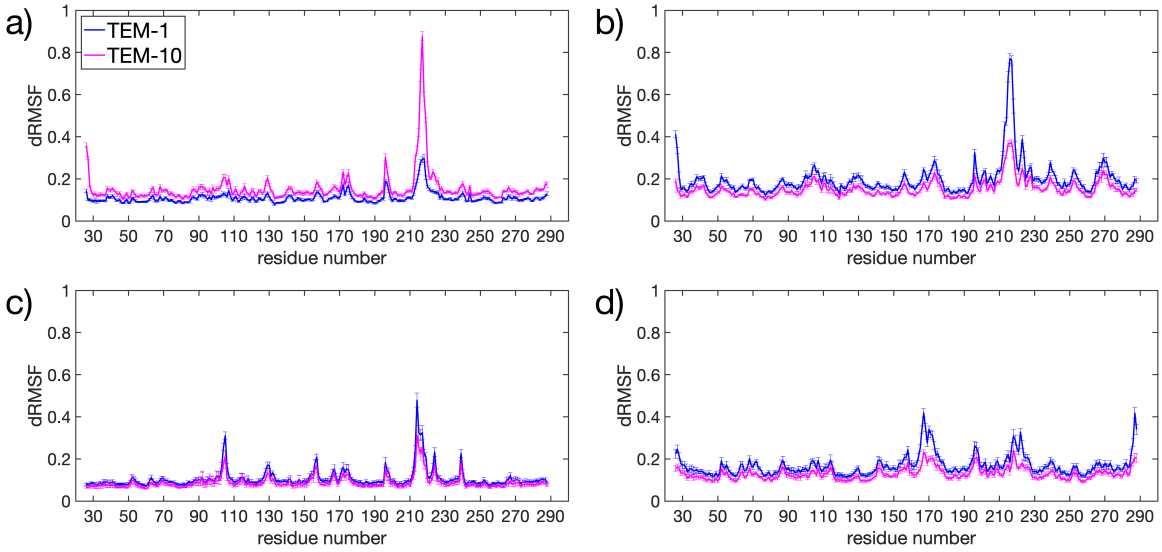

Figure 4.9: Changes in dynamics between TEM-1 and TEM-10 while in complex with a) AIC, b) AXL, c) CEF, d) CAZ ligands.

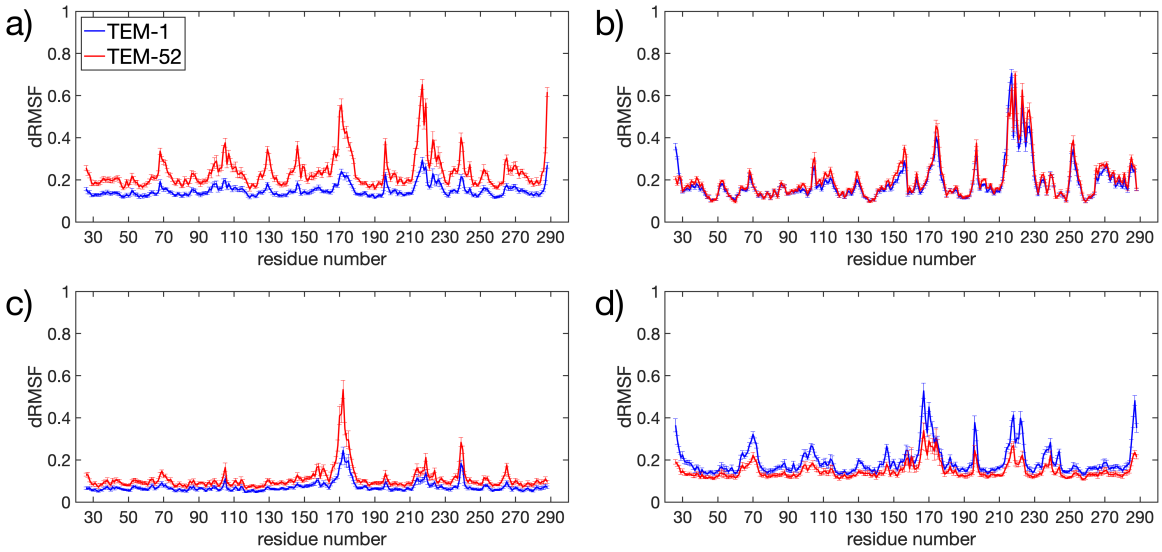

Figure 4.10: Changes in dynamics between TEM-1 and TEM-52 while in complex with a) AIC, b) AXL, c) CEF, d) CAZ ligands.

All 16 of the holo systems, with 8 replicate simulations per system, were compared head-to-head to elucidate ligand-specific changes in dynamics due to mutations. For each ligand, the different beta-lactamase mutants were compared similar to how the apo enzymes were compared in Section 4.4.2. Specifically, for each ligand, TEM-1 was compared to TEM-2 (Figure 4.8), TEM-10 (Figure 4.9), and TEM-52 (Figure 4.10). These comparisons reveal how mutations change beta-lactamase dynamics to accommodate different ligands.

In the case of TEM-2, ligand binding induced changes to the enzyme motions primarily centered at the H10-H11 loop. This region showed the greatest spike in discriminant motion for AIC or AXL binidng. At this loop, TEM-2 had higher dRMSF compared to TEM-1 when bound to AIC, whereas TEM-1 had slightly higher dRMSF than TEM-2 when bound to AXL. This difference is interesting as TEM-1 and TEM-2 have similar affinities for AIC and AXL, thus changes in substrate specificity do not account for this. Visualizing the simulations in PyMol revealed notable insights. For AXL binding, some TEM-1 simulations exhibited significant conformational change in the H10-H11 loop compared to the apo crystal structures. In the simulation started from the 3JYI crystal structure specifically, the loop assumed an extended conformation that is flipped away from the protein and reaching into the solvent. For AIC binding, no such dramatic conformational changes could be found in the TEM-2 simulations. However, there was a considerable amount of motion around the native conformation of the loop, which could explain the spike in Figure 4.8 a. For CEF binding the enzymes showed little differentiating motion, while for CAZ binding, some motions seem to be lost in TEM-2 at the omega and H10-H11 loops. Notably, neither TEM-1 or TEM-2 bind well with CEF or CAZ.

TEM-10 exhibits a similar response to ligand interactions compared to TEM-2, despite being an extended-spectrum enzyme. Interestingly, TEM-10 retains catalytic efficacy for AIC experimentally, suggesting that the dynamics that facilitate bind-

ing to these drugs must still be present in TEM-10, even with its mutations. Unlike TEM-2, TEM-10 does show dynamic shifts compared to TEM-1 upon binding CEF, particularly around the H3-H4 and H10-H11 loops. The similarity in dynamic shifts for TEM-10 and TEM-2 in complex with AIC, AXL, or CAZ, despite the enzymes having different substrate specificity profiles, is striking, and more work may be needed to confirm if these dynamic changes actually impact substrate specificity in beta-lactamase or are artifacts of the molecular dynamics simulations.

Finally, TEM-52 shows the most pronounced dynamic shift among the mutants for all ligands. The dRMSF highlights significant changes in dynamics that occur at the omega loop and H10-H11 loop which is consistent with the rest of the analyses in this chapter. These results reveal the unsurprising reality that both the mutations and ligand identity have important roles in changing the motions the TEM beta-lactamase.

4.6 Using Functional Dynamics To Predict Extended Spectrum Resistance

An important result of this work is that each beta-lactamase system studied exhibits its own dynamic signatures. This raises the question of whether these signatures can be used as predictors of enzyme kinetics. As discussed in Section 3.4.2, SPLOC can can use the features it finds to predict whether or not data comes from a particular class by computing a discovery likelihood. In this section, the feasibility of using functional dynamics in beta-lactamase to predict substrate specificity profiles is investigated.

Experimental data on the minimum inhibitory concentration (MIC) for each of the antibiotics used in this study against each of the enzymes was collected from literature and is shown in Table 4.5. This table provides valuable information about the function of enzymes that provides SPLOC with a way to partition training sets. From the table, it can be seen that the full complexity of "function" in beta-lactamase is not adequately described by simply labelling an enzyme "wild type" or "extended-

Table 4.5: MIC values collected from literature sources. MIC measures how high of a concentration of a drug is needed to inhibit the growth of a bacteria. Low MIC values indicate the antibiotic is not susceptible to beta-lactamase resistance, while high MIC values indicate that it is resisted. Ligand types are denoted on the table, and TEM-1 and TEM-2 represent wild-type enzymes while TEM-10 and TEM-52 represent ESBL-type enzymes.

MIC ($\mu g/ml$)	TEM-1	TEM-2	TEM-10	TEM-52
AIC	> 128 [180]	-	> 256 [138]	256[181]
AXL	> 1024 [139]	-	-	> 1024 [139]
CEF	< 0.125 [180, 182]	0.125 [182]	< 1 [138, 182]	32 [139]
CAZ	< 0.25 [180, 182]	0.5 [182]	< 64 [138, 182]	128 [139]

spectrum".

For example, TEM-10 is considered an extended-spectrum enzyme because CAZ has a high MIC value for bacteria which produce this strain of beta-lactamase. However, TEM-10 does not hydrolyze cefotaxime as efficiently. It should be noted that MIC values for TEM-2 against AIC and AXL could not be found, however literature supports that it has the same catalytic properties as TEM-1. [180] All of the enzymes used here appear to be able to effectively resist ampicillin and amoxicillin, with the exception of TEM-10/AXL for which no experimental data could be found. As such a meaningful comparison of enzymes which can resist broad-spectrum antibiotic and those that cannot is not able to be made.

On the other hand, it is possible to divide the enzymes into two groups based on whether or not they are able to efficiently resist against extended spectrum cephalosporins CEF and CAZ. In this comparison, TEM-1 and TEM-2 in complex with CEF and CAZ are considered functional as beta-lactamase is unable to effectively resist against the antibiotics. TEM-52 in complex with CEF and CAZ are considered non-functional. Despite being considered an extended spectrum beta-lactamase because it can resist ceftazidime, TEM-10 has split function between CEF and CAZ. As it does not fall cleanly into either category, it will not be included in the training set.

Holo simulations are used to train SPLOC, as these capture the enzyme motions

which have been induced or quenched by the antibiotic's presence. Bootstrapping was employed to construct data packets in the same way as in the previous section, however the number of bootstrapped data packets used was lowered to 5 rather than the 15. Using raw trajectories as data packets was also tried, but it did not yield d-modes to classify with. SPLOC was run for 10 replicate trials, in its neutral MO, and using an identity matrix as the starting basis set.

The bootstrapped data packets were generated using with exactly 8000 samples each, the same number of samples in each trajectory. After SPLOC was trained the resulting discriminant mode basis sets from each replicate trial were used as classifying vectors to predict the discovery likelihood that apo simulations of the four mutants would not bind to extended-spectrum antibiotics. The discovery likelihoods are reported as an average over the predictions from all basis sets. Because the statistics were the same, both the raw trajectories and bootstrapped trajectories from apo simulations were classified using SPLOC. Classifying the bootstrapped trajectories resulted in better results, and so only results for the bootstrapped trajectories are shown.

The resulting classifier, shown in Figure 4.11 a, shows that the likelihood that any of the 4 beta-lactamase enzymes will not bind with extended-spectrum antibiotics is high. Despite this, there is a noticeable drop in likelihood for TEM-52, a known binder of both CEF and CAZ. This suggests the classifier detected that TEM-52 is more likely to bind with these antibiotics compared to TEM-1 and TEM-2. TEM-10, which is observed to have partial extended-spectrum function, appears to have a higher likelihood not to bind to extended-spectrum antibiotics compared to TEM-52, which appears consistent with experimental results in Table 4.5. The table indicates that TEM-52 binds well with CEF and CAZ, while TEM-10 only binds well with CAZ. It is possible that with increased sampling, either through further MD simulation or improved bootstrapping, the relative likelihood contrast can be improved to better

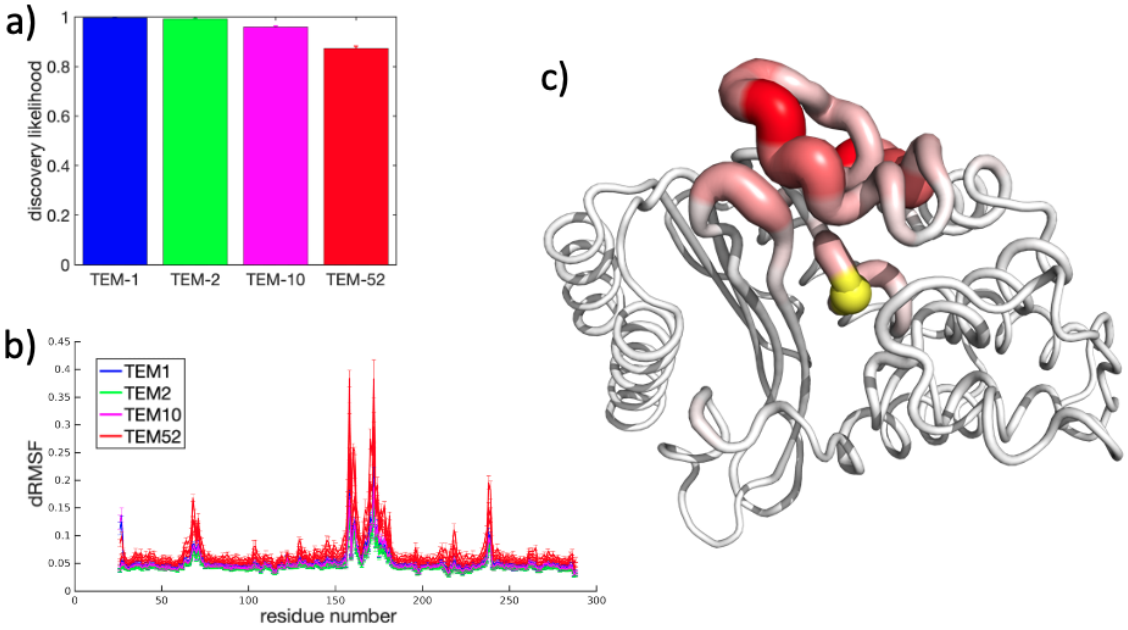

Figure 4.11: Panels a-b) show SPLOC results for comparing penicillin-like holo dynamics. a) shows the discovery likelihood that an enzyme is not going to bind well to an extended-spectrum enzyme and b) shows where the dynamic shifts occured during holo simulations which allowed SPLOC to classify the apo systems. c) shows the dRMSF for TEM-52 projected onto a beta-lactamase structure. dRMSF values under 0.07 were thesholded to 0.0 in pymol, and the yellow sphere represents Ser70.

reflect experimental resistance profiles.

In Figure 4.11 b-c, the dRMSF from this SPLOC comparison is shown, representing where on beta-lactamase the dynamics that formed the basis of the classifier were found to occur. Primarily these motions occur at the omega loop, particularly at the top most part of the loop and just after the C-terminal of the loop. Additionally, there is a smaller peak centered near Ser70 (shown as a yellow sphere in panel c) and at the beta turn near the omega loop.

With the relatively few examples of functional and non-functional proteins used in this comparison, the results in the section should not be considered definitive, but a proof-of-concept for how functional dynamics can be used in a practical application to predict ligand binding properties of enzymes. To test this more thoroughly, more enzymes with different binding qualities, especially with broad-spectrum antibiotics, should be simulated and tested. Moreover, if enough examples of proteins that do and do not bind with each antibiotic are simulated, then a per-antibiotic classifier could be constructed. With the addition of more observations of different proteins, it is a possibility is that the results found here are statistically trivial and get washed away in noise. However, validation of this methodology could present a novel approach to identifying antibiotic resistance in unknown beta-lactamase enzymes.

4.7 Discussion and Future Directions

In this chapter the functional dynamics of four TEM beta-lactamase were thoroughly characterized using MD simulations combined with SPLOC, a novel discriminant analysis method. Pairwise comparisons of simulations using different groupings were able to decouple statistically different motions from the global conformational motions between mutant enzymes in a variety of different conditions. These dynamic changes elucidate the role dynamics play in acquiring resistance to new antibiotics. Although the differences observed in this work have revealed important mechanisms of substrate recognition and binding in beta-lactamase, there is still more left to un-

derstand. In this section, some of the key findings about beta-lactamase functional dynamics are summarized and several future extensions and modifications to the methodology are proposed.

4.7.1 Hot-spots For Dynamic Change In TEM Beta-Lactamase

Throughout the prior sections, multiple loci on the beta-lactamase enzyme were identified as sites where changes to the dynamics of the protein are connected to changes in function, either via mutation or ligand binding. Particularly these regions most often fell either at important catalytic residues or on loops that surround the binding pocket.

The omega loop (residues 163-178), a feature common to all class A beta-lactamase, exhibited the most frequent dynamic changes. This loop has been previously postulated to be important for both substrate recognition and catalysis. [76] Notably, Glu166, a residue on this loop, is proposed to directly take part in the hydrolysis of beta lactam rings. The results shown here supports that this loop is critical for protein function.

In section 4.4.2 it was shown that the omega loop was strongly impacted by mutations that lead to extended-spectrum TEM-10 and TEM-52 beta-lactamase. The change in motions was most prominent in TEM-10, likely because one of the mutations (R164S) is on the loop itself. However, the dRMSF for the ESBL mutants at the omega loop was much higher compared to TEM-1.

These findings suggest that the mutations induced specific motions in the omega loop, allowing the enzyme to accommodate the larger extended-spectrum antibiotics. Of the ligands used here, the data suggests that CAZ recognition is most sensitive to motion in the omega loop. Figures 4.8-4.10 illustrate that each mutant bound to CAZ (panel d) had lower dRMSF compared to TEM-1, implying that the motion induced into the enzymes by mutations was dampened upon binding with the ligand. In contrast, changes in omega loop motions compared to TEM-1 due to CEF binding

was only observed in TEM-52.

On the opposite side of the binding pocket, the H10-H11 loop, along with most of helix 11, appears to be another region on beta lactamase which is susceptible to dynamic changes. When comparing apo simulations, only TEM-52 exhibited a dynamic shift here. This indicates that changes in the dynamics of the H10-H11 loop are less connected to beta-lactamase mutations, rather the locus to be more impacted by ligand binding. In Figure 4.6 c and d, this loop showed the strongest change in dRMSF for TEM-10 and TEM-52 when comparing how the enzymes react to ligands. This response appears was observed in both broad-spectrum and extended-spectrum antibiotic interactions with beta-lactamase.

In Section 4.5 the dynamics of this loop were further probed using the expanded L-library, revealing that the H10-H11 loop responds to binding by all ligands in all 4 mutants. Each system, when compared to TEM-1, had a somewhat unique dRMSF pattern. However, for enzymes with different mutations bound to the same ligand, the dynamic signatures all had spikes at the H10-H11 loop.

Based on these observations and its proximity to the active site, these results suggest that motion in the H10-H11 loop facilitates conformational changes of the loop, allowing it to better accommodate different ligands. A potential role is for this loop is helping to stabilize the ligand during hydrolysis. In order to interact with small AIC and AMX molecules, the H10-H11 loop would be required to have extra conformational flexibility to reach into the binding pocket where the ligands reside. Indeed, this loop was observed to take a wide range of conformations during the MD simulations. Some representative conformations are shown in Figure 4.12. Alternatively, CEF and CAZ are very bulky and the loop would not have to reach as far. During the simulations, with CEF in particular, the ligand tended to stretch itself across the binding pocket and anchoring its ends near the omega loop and the H10-H11 loop. These observations provide additional support for the loop's proposed

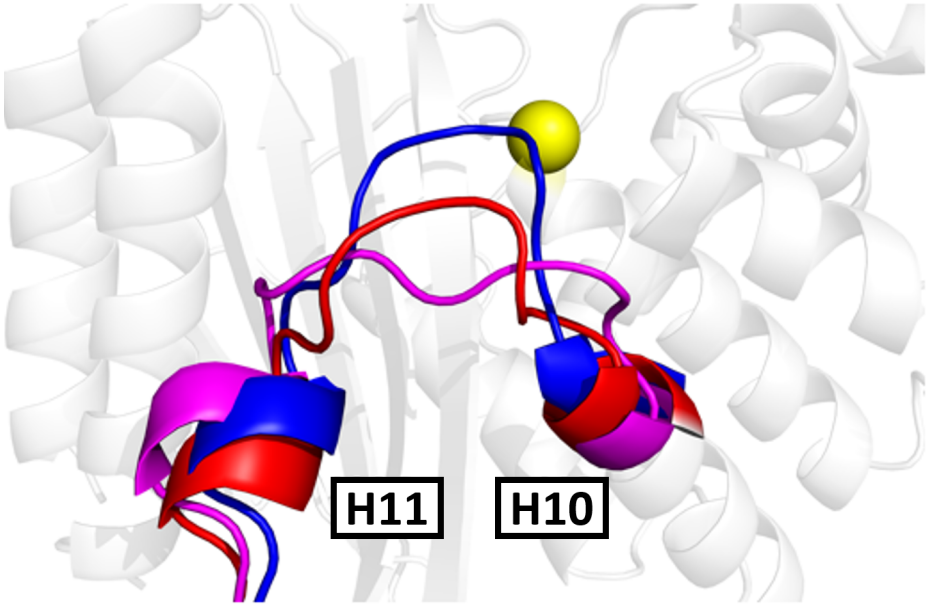

Figure 4.12: Selected conformations of the H10-H11 loop. The location of Ser70 is shown for reference in yellow. Figure Credit: [1]

role in stabilizing the enzyme/antibiotic complex. Notably, it has been suggested in literature that Val216 on this loop, along with Arg244, plays role in stabilizing the catalytic water needed to donate a proton for the hydrolysis process. [145, 169]

Lastly, another region which has a propensity for changing its dynamics was the loop connecting helix 9 and 10. This loop is on the complete opposite side of the protein compared to the binding site. This distant loop acts as a hinge connecting the omega loop to the H10-H11 loop and has been suggested to be important for propagating allosteric signals in beta-lactamase [179], however there is limited prior evidence in the literature to suggest that it directly impacts the catalytic function of the protein.

In Section 4.4.2, a major shift was observed in this loop when the protein underwent mutation. In Section 4.4.3, this loop also exhibited changes in dynamics when the apo motions of the protein were compared holo motions. Subsequent comparisons between different holo systems demonstrated a weaker dRMSF signal at this loop. Based on these findings, the H9-H10 could be crucial for alleviating stress imposed on

the beta-lactamase during changes such as mutations or ligand binding. Not only did it undergo dynamic changes due to both mutations and ligand binding, but it also connects two other interesting regions on beta-lactamase that are shown to undergo changes in dynamics as well.

4.7.2 The Nature Of Functional Dynamics In TEM Beta-Lactamase

Most of the dynamic changes discussed in this work involve motions that are small in magnitude compared to the global motions of the protein. dRMSFs values typically ranged from 0.1-1.0 Å, whereas the total RMSF reaches up to 3.0 Å. Also, dRMSF motions tend to be localized at specific regions of the enzyme structure, rather than uniformly distributed.

Together, these observations reveal that the nature of functional dynamics in TEM beta-lactamase are primarily involved small, localized motions. These motions likely could not be identified by direct observation of the MD trajectories and weren't captured the essential dynamics of beta-lactamase in Section 4.2. Consequently, the unbiased and data driven hypothesis formation of SPLOC was essential for elucidating these motions.

To demonstrate the differences between "important motions" found by SPLOC and essential dynamics, the discriminant modes found in Section 4.4.2 were compared to PCA modes found using the same data. Figure 4.13 shows the cumulative overlap between the top N PCA modes, sweeping from $N = 1$ to the size of the system (789), using the discriminant, indifferent, undetermined, and full subspaces from SPLOC. The motions found in SPLOC d-modes do not accumulate overlap with the PCA modes until the lowest variance PCA modes were added, at around 400 PCA modes. In contrast, the cumulative overlap with i-modes increases almost linearly. This result does not mean that essential dynamics is unable to find functional dynamics in general, however, it does illustrate why it is important to consider the assumptions underlying a model before using it. The assumption in PCA is that the most variant

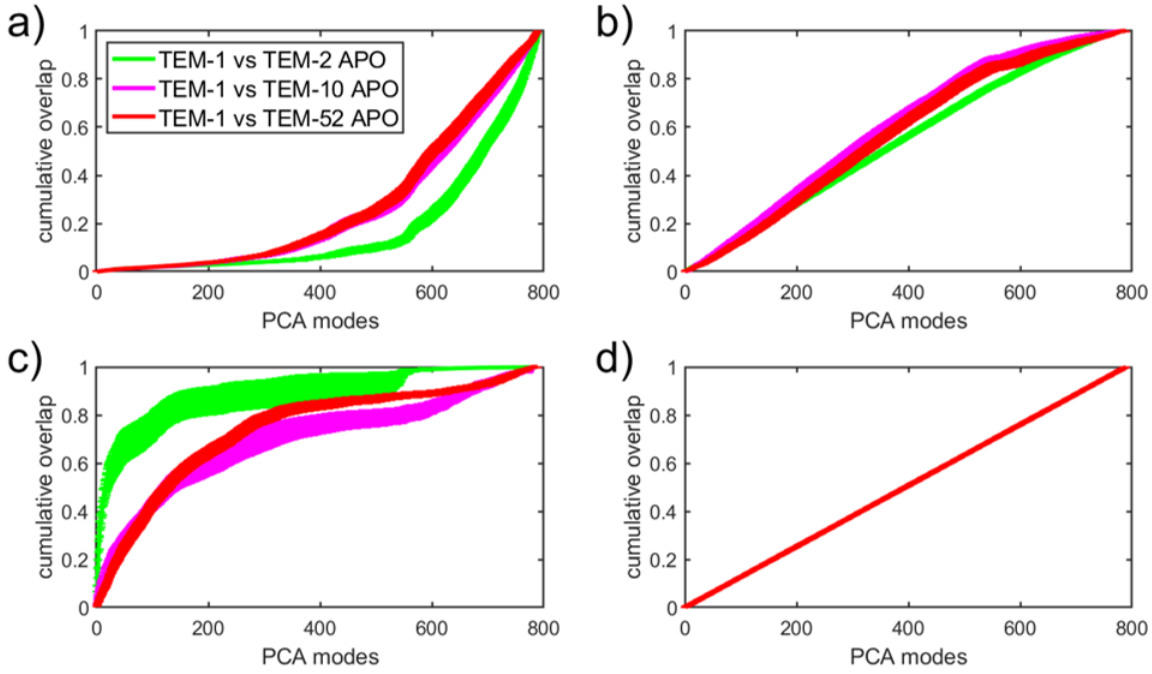

Figure 4.13: The average cumulative overlap (CO) between SPLOC modes and PCA modes in a) discriminant space, b) indifferent space, c) undetermined space, and d) over the entire basis set. CO is computed for each mode in the SPLOC subspace by summing over the top n PCA vectors. The results here average over all replicate SPLOC runs, and all modes within each subspace, with error bars representing standard error. Figure Credit: [1]

motions are the most important. As the functional motions in beta-lactamase are not the most variant motions of the enzyme, PCA is unable to capture them.

4.7.3 Long Time Scale Dynamics

Protein motions exist at a wide range of timescales, yet molecular dynamics simulation is limited to the picosecond to millisecond range. In this work, simulations were conducted for 500 nanoseconds at a time, and to effectively increase this limit, a shotgun approach of running multiple replicate simulations with different starting structures was employed. While the crystal structures available in the PDB and the results of the simulations presented in Section 4.2 suggest that TEM beta-lactamases are generally rigid and do not undergo functional global conformation changes, the possibility of large conformational changes beyond the scope of what was simulated

here cannot be completely ruled out.

Computational and NMR studies in the literature suggest that the omega loop is capable such larger motions [79, 183], albeit in the millisecond time domain. Even within the 160 simulation library used here, there were examples of the omega loop pushing itself from the main body of the protein and extending into the solvent, hinting at the existence of such motions. To understand if these motions represent true functional dynamics of beta-lactamase or if they are due to random thermal fluctuations, or an insufficiently parameterized forcefield, longer simulations are needed to obtain better sampling. One attractive approach to this is by using accelerated MD methods.

To this end, some work has begun exploring the use of coarse-grained forcefields, particularly the MARTINI forcefield which has been used in many accelerated MD studies of proteins. [184, 29] In MARTINI, groups of about 4 or so atoms in an all-atom structure of a protein are represented as beads. Each residue on the protein contains a backbone bead and 1-4 side-chain beads, depending on the structure and local environment of the amino acid. Common waters, ligands, cofactors, and ions also have corresponding MARTINI coarse-grained representations. Each bead is then simulated as a coarse grained atom. Because multiple atoms are represented by a single object, all-atom forcefield parameters cannot be used, and MARTINI has specialized forcefield parameters for each potential type of bead. [185]

In addition to reducing the number atoms needed in a simulation, the MARTINI forcefield also allows for the use of larger time steps, up to 20 femtoseconds, compared to the 2 femtoseconds required for stability in all-atom simulations. This leads to a massive speedup in simulation time.

Initial testing of this forcefield with beta-lactamase allowed 500 nanoseconds of dynamics data to be collected in 1.42 hours, which is a fraction of the more than 48 hour wall time required for a single all-atom simulation of beta-lactamase at the

same timescale. Despite the impressive wall time speed up, further optimization of the simulation parameters are needed to maintain enzyme stability before meaningful molecular dynamics can be simulated. During the test simulation, the protein immediately unfolded, something that was completely unobserved in the all-atom simulations, and a hallmark of an improperly parameterized simulation.

4.7.4 Outlier Processing

During the course of this research, SPLOC was not able to find discriminant motions in beta-lactamase without using a bootstrapping approach to construct data packets. Bootstrapping provided several benefits, including increasing the OPV in each data packet and allowing each data packet to characterize the variation across multiple independent simulations of the same class. This greatly decreased the variance between data packets in SPLOC training sets, increasing the likelihood for SPLOC to converge.

It is well known that molecular dynamics simulations are stochastic, as is the nature of molecular dynamics. [186] Unless an random number generator seed is used, the variance between any two simulations of the same system may be large, and some motions may be considered to be outliers. Here, an outlier can either represent a statistical fluctuation of the protein, or it could represent a rare event which is important for enzyme function. The challenge in interpreting molecular dynamics is to identify which is which. It should be noted that, in MD, the term outlier is subjective to the timescale used. A large fluctuation of the protein at 500 ns may represent statistical noise at 500 ms. [187]

Due to the bootstrapping method employed in this work, the effect of outlier motions, which may only occur in 1 MD trajectory, will be amplified because the motions in each simulation is equally represented in the bootstrapped data packet. To address this, a method of preprocessing which uses SPLOC was conceived for use in future work. This method removes discriminant motions prior to comparing trajectories

with SPLOC.

Consider a set of N MD trajectories for a particular system, denoted by $S = \{X_i\}_N$. For each trajectory, the simulation frames are randomly partitioned into two samples, S_1 and S_2. SPLOC is then used to compare the all the trajectories from one partitioning to the other, resulting in a set of basis vectors, U, that describes similarities and differences in the simulations coming from the same system. The i-modes, represented by $\{|k_I\rangle\}_{S_I}$, capture the motions that are the same between each trajectory in the system. The matrix of indifferent modes representing the indifferent motions between simulations of class 1 and class 2 are denoted as U_1^i and U_2^i respectively.

The original trajectories are projected onto both sets i-modes, resulting in a new set of generalized coordinates to represent the system. If class 1 yields $N1_I$ i-modes and class 2 yields $N2_I$, then a total of $N_I = N1_I + N2_I$ new independent variables can be used to probe the systems. Each trajectory, projected onto each of the N_I coordinates can be SPLOCed to find differences.

The resulting basis vectors in the discriminant space will be N_I dimensional. Each basis vector can be split into two portions: the first $N1_I$ elements represent differences in the indifferent motions coming from class 1, and the last $N2_I$ elements represent the portion of the vector representing differences in the indifferent motions coming from class 2. The matrix of vectors can be divided into these portions, a set of partial vectors V_1^d and V_2^d respectively. Similarly the i-modes can be split into contributions from each class, V_1^i and V_2^i. These vectors represent the discriminant and indifferent motions in the basis set representing the indifferent motions of the systems themselves.

Using the properties of linear algebra, each discriminant mode can be "un-projected" back into the original basis set, representing the atomic coordinates of the trajectory. The trajectory-basis d-modes ($D = \{|k\rangle_d\}$) and i-modes ($I = \{|k\rangle_i\}$) that represent dynamics can be computed as in Equation 4.1 and 4.2, where the basis vectors are

represented by the column space of the matrix.

$$D = U_1^i * V_1^d + U_2^i * V_2^d \tag{4.1}$$

$$I = U_1^i * V_1^i + U_2^i * V_2^i \tag{4.2}$$

The benefit of using a complicated scheme like the one described here is that, in the absence of outlier motions, each data packet used in SPLOC can represent a single trajectory. This closer reflects the true dynamics of the system and avoids the potential over-representation of fluctuations between MD simulations. This method was fully prototyped and tested using a subset of the beta-lactamase dataset. When the outlier detection was applied, d-modes could be found using the filtered trajectories as data packets. However, SPLOC could only find inconsistent or weak d-modes, resulting in a noisy and difficult to interpret dRMSF. Further work is needed to refine the outlier detection before it can be readily utilized. Note that it is up to the user to determine if applying an outlier detection before using SPLOC to identify functional dynamics is appropriate. If the functional dynamics of a system are random fluctuations, then using outlier detection will throw out any interesting signals.

4.8 Conclusion

Using SPLOC, the functional motions of TEM beta-lactamase have been thoroughly characterized. SPLOC found that functional motions that facilitate ligand binding and recognition are often found to be small localized motions in loop regions in or bordering the active site. Some distal sites on the protein were also found to show a change in motion due to mutation or ligand binding, including regions on beta-lactamase identified as potential hot-spots for dynamic allostery.

TEM beta-lactamase are able to bind to a variety of classes of antibiotics, and in this work the specific motions which facilitate binding to ampicillin, amoxicillin, cefo-

taxime, and ceftazidime with TEM-1, TEM-2, TEM-10, and TEM-52 beta-lactamase were analyzed. Each enzyme/ligand pair was found to have its own unique dynamic signature characterizing how mutations and ligand binding impacts the dynamics of the protein.

Finally, the connection between functional dynamics and function was tested by considering whether the dynamic signatures identified by SPLOC could be used to predict whether or not beta-lactamase enzymes would exhibit extended-spectrum antibiotic binding. The resulting classifier did show a difference in likelihood between broad-spectrum and extended-spectrum enzymes, which suggests that beta-lactamase motion is fundamentally related to substrate specificity. While this type of analysis is not practical for real time enzyme function classification, it represents a novel perspective for understanding how antibiotic resistance arises in novel beta-lactamases, and in general deepens the connection between protein dynamics and protein function.

CHAPTER 5: PEPTIDE INHIBITION OF ANTIBIOTIC RESISTANCE

5.1 Motivation

In this chapter, the feasibility of designing de-novo peptide inhibitors for beta-lactamase is investigated using a novel peptide design approach called pepStream. PepStream was originally developed as part of another project to determine peptides that can bind to disordered regions on proteins. In this work, the method is adapted and tested on TEM-1 beta-lactamase, a well-ordered globular protein.

Beta-lactamase inhibitors have revitalized the efficacy of beta-lactams such as amoxicillin. [23] Current beta-lactamase inhibitors, including Clavulanic Acid, Sulbactam, and Tazobactam, target the active site of beta-lactamase. These inhibitors work by either by forming a stable Michaelis-complex with the enzyme using their own beta-lactam ring motifs, or take residence in the active site pocket, occluding antibiotics from reaching the catalytic SER70.

Unfortunately, the emergence of inhibitor resistant beta-lactamase poses a prominent threat to global health. [188] In Chapter 4, the dynamics of four TEM beta-lactamase were analyzed, and motions with importance to substrate recognition and binding were identified. These dynamic motions occurred both near and far from the catalytic site on the enzyme, suggesting that modulating the dynamics in these regions of beta-lactamase can control substrate recognition.

Here, the ideas from Chapter 4 are expanded upon, and pepStream is employed to predict what kind of peptides might preferentially bind to the regions, where TEM-1 exhibits functional dynamics. Ideally, the peptides predicted by pepStream will bind to the enzyme and disrupt the motions needed for antibiotic recognition. These peptides represent potential de-novo beta-lactamase inhibitors, which use a novel

inhibition mechanism of targeting the functional dynamics of the protein.

5.2 Automated Peptide Design with PepStream

PepStream is unique among other peptide design platforms as it combines sequence-based and structure-based approaches to optimize binding peptide sequences. [189, 190, 191] The scope of this chapter is not the conception or design of the pipeline, rather an new implementation into a single, easy to run SLURM-script-based pipeline. Parts of the pipeline have been upgraded to include state-of-the-art tools in structural bioinformatics and methodological improvements, which allow for higher quality binding partners to be found. These upgrades have made pepStream more accessible and computationally efficient.

Prior to this work, pepStream had been implemented in a series of bash and python scripts, with the basic workflow described as follows. There are three major stages to running pepStream: sequence diversification, peptide construction, and docking-based filtering. In the sequence diversification stage, sequences are defined representing the protein of interest, and the subsequence of the protein representing the binding target. The target sequence, referred to here as a Molecular Recognition Feature or MoRF, terminology borrowed from intrinsically disordered proteins [12], is sliced into footprint sequences by successive binary splitting.

These footprints are run through PSI-BLAST [192, 193] to find homologically distant matches in other proteins, generating a position-specific scoring matrix (PSSM). The PSSM is used to computationally mutate the footprint sequences to create a list of "seed sequences". Using BLASTp [192], the PDB is searched for proteins with homologous regions that match the seed sequences and have a corresponding solved structure.

In the peptide construction stage, the structures obtained from the BLASTp search are mined to look for fragments of structures which interact with the region on the protein homologous to the target sequence. The underlying concept is to leverage

interactions that nature has optimized through evolution to design peptides, rather than guessing sequences randomly. Using the PDB structures, the amino acids which are found to interact with the matching sequence are strung together into a contiguous peptide fragment, called a cMoRF (complementary MoRF). cMoRFs are potential candidates for being a peptide binder for the protein. Finally, the structure of the full protein and each cMoRF peptide are predicted.

In the last stage of the original pipeline, cMoRFs are docked into the protein in several steps to filter out peptides which do not bind well. Firstly, all of the peptides are globally docked to the protein with no restraints to determine the top $X\%$ of cMoRFs that interact strongly with the protein. Next, the strongest interacting peptides are docked again, however this time the docking area is restrained to the target sequence. The peptides with the highest docking scores from this second docking phase were considered the best binding peptides in the original pepStream pipeline.

Specificity can be evaluated for the peptides by docking the top-ranking peptides to alternate sites on the protein and comparing the relative docking scores. If a peptide has the highest docking score for the target sequence, then it is considered to have specificity for this region. The final output of pepStream was a list of peptide candidates that are predicted to strongly bind with the protein of interest, with specificity for the target region.

The original pepStream pipeline was designed to be modular, specifically the structure prediction and docking parts of the pipeline, allowing for the substitution of different software methods as needed. In this version structure prediction was performed with i-Tasser, and docking was performed with Z-Dock. [194, 129]

In the following sections, a re-implementation of pepStream will be described, followed by the results of using pepStream to predict peptides which bind to various sites on TEM-1 beta-lactamase.

5.3 Implementation and Updates to pepStream

For this work, pepStream was implemented as a self-contained bash script, accompanied by a library of helper bash/python scripts which are called by the main script. The script was designed to run on a GPU-enabled High-Performance Computing (HPC) Cluster, with resource usage optimized for UNCC's Leo cluster. This cluster is equipped with 8 Nvidia A100 GPUs and 128 Intel Xeon Cores. Each pepStream job utilizes 1 GPU and 16 CPU cores, allowing 8 jobs to be run concurrently.

The user interface for pepStream was simplified to a single parameter file which is submitted to pepStream at the command line. All parameters for a job are defined in this input file, where users simply have to fill in the blanks with their system of interest, target, alternate target binding sequences, and several other parameters related to each of the modules in the pipeline.

All outputs for pepStream are organized into a single specified directory designed for easy access and organization of relevant data. Currently, most of the data is not compressed, and each run may generate dozens to hundreds of gigabytes of data depending on how many cMoRFs are identified. This version of pepStream takes a few days to run from start to end, again dependent on how many cMoRFs are identified. However, a checkpoint system was implemented so that pepStream can be restarted if the job crashes or runs out of walltime.

5.3.1 Updates to Structure Prediction

In the updated version of pepStream, i-Tasser was replaced with OmegaFold [195], resulting in a significant speed up in the structure prediction phase. With OmegaFold, the prediction of a single structure can take less than a minute. The primary drawback of using OmegaFold is that it only produces a single model of a protein structure, whereas i-Tasser was capable of producing multiple models. This drawback is significant because rigid docking methods are currently being used in pepStream, and for

these models conformational shape is an important factor in proper receptor-ligand docking.

OmegaFold can be forced to produce several different models by adjusting some of the parameters. However, this may decrease the quality of each model. To address this, a more sophisticated methodology, called the conformation generation pipeline (CGP), was devised to produce multiple quality conformations of a structure.

The CGP is run in three main steps. First, OmegaFold is used to generate an initial conformation of the protein. This structure is then run through a geometric simulation using FRODA, which is part of the FIRST software. [196, 197] A Python script then analyzes this trajectory to find the most diverse conformations sampled by the simulation. These diverse conformations are found by computing the principle components of the FRODA simulation and applying a clustering algorithm to find a diverse set of $NCONFS$ conformations. To obtain physically valid conformations, the most diverse conformations, including the initial OmegaFold structure, are energy minimized in explicit water using GROMACS.

Currently many of the parameters for FRODA and GROMACS are hidden from users. However, it would be easy to extend the pepStream input parameter file to include them in a subsequent update to the pipeline. FRODA is run with a hydrogen bond energy cutoff of -2 and generates a total of 100000 simulation steps, outputting a structure every 10 steps to yield a final trajectory of 10000 structures. For GROMACS, the structure is solvated in a cubic box, the system neutralized with ions, and the whole system is minimized using steepest descents until the maximum force on the system is less than 500 kJ/mol.

Conformational diversification is performed because the rigid-rigid docking method in the next step of the pipeline is highly dependent on the conformation of both the receptor and ligand. Docking is performed pairwise between $NPROT$ diversified conformations of the receptor protein and $NPEP$ diversified conformations of the

peptide, rather than just between one structure of each. This accommodates some of the intrinsic flexibility within protein-peptide complexes, while not sacrificing too much computational complexity.

OmegaFold was chosen over the more sophisticated AF2 model, due to its ability to predict structures of comparable quality, while taking a fraction of the time. On the UNCC cluster, AF2 required around 40-50 minutes to predict the structure of TEM-1 beta-lactamase, and around 20-30 minutes for a short peptide. Using the full CGP, including OmegaFold, FRODA, and GROMACS, this time was reduced to about 20 minutes for beta-lactamase, and between 1 and 2 minutes for each peptide. OmegaFold achieves its speed increase by bypassing the computationally expensive MSA construction, which is the bottleneck step in AF2.

5.3.2 Updates to Docking Methodologies

The docking program in pepStream was updated from Z-Dock [129] to MEGADOCK [128], primarily to increase the speed of the program. MEGADOCK, having been developed by the same group, is the successor to Z-Dock. PepStream performs docking for all pairwise combinations of protein and peptide structures, generating 1000 poses per complex. Even using ultrafast GPU-accelerated calculations, this step remains pepStream's most rate-limiting step.

In the original version pepStream, a two-phased approached to docking was employed, unrestrained global docking followed by restrained target docking. Only the top percentage of peptides from the global were selected to move to the targeted docking phase. Through analyzing the results of several tests, it was determined that the MEGADOCK scoring function had a moderate to strong dependence on the length of the peptide sequences. This dependence is less apparent when docking to specific regions. This can be seen in Figure 5.1.

Additionally, a weakness in the two-phase docking approach in the original pepStream was realized. Performing global docking first followed by targeted docking of

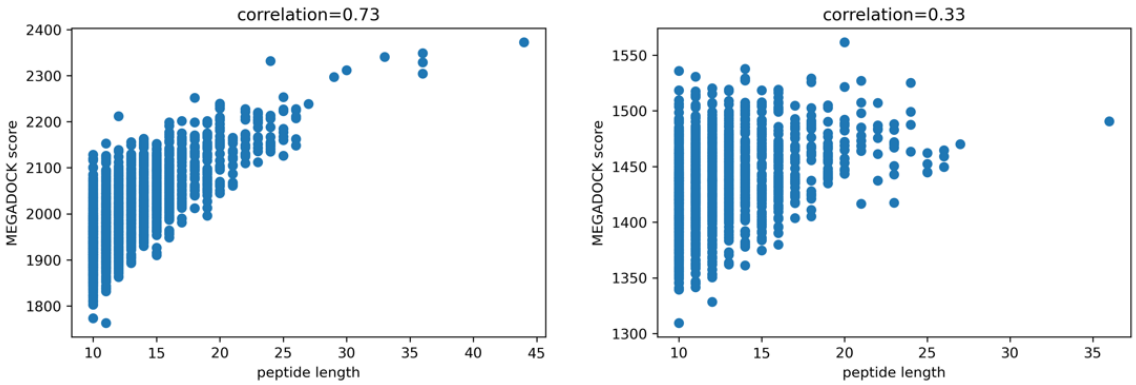

Figure 5.1: Average docking scores for a) global and b) targeted docking in MEGADOCK.

only the top $X\%$ of cMoRFs can potentially lead to the omission of peptides that bind well to the target. This situation occurs when the highest docking scores for peptides binding to the target region are lower than the highest docking scores for the peptides that don't bind at the target region. Only in the case where the two sets of high scoring peptides overlap will pepStream be able to identify the correct high affinity binders for the target.

As an example, a protein may exhibit high affinity scores for peptides near its active site, where it has been evolved to bind ligands. However, if a peptide binding far from the active site were wanted, then pepStream would generate false positive sequences of peptides that preferentially bound at the active site. Another example of when this might happen is when a protein has multiple binding sites. In such cases, peptides that interact strongly with one binding site may overshadow those that interact at the other.

For these reasons, a modified docking procedure that eliminates global docking completely is proposed. In this procedure, the first phase of docking uses a docking area restrained to the target region to find the peptides which have the strongest interaction with the target. Only the peptides which interact with the target move on. The second phase performs restrained docking where the docking area is restrained to a number of user-supplied alternative docking sites to test for specificity. The high

affinity peptides can then be reranked by which have the highest specificity.

In this scheme, the resulting peptides will be those with the highest affinity and specificity for binding at the target, rather than the peptides which interact most strongly anywhere on protein. Additionally, removing the global docking reduced the impact of the MEGADOCK sequence length-dependent scoring function on the predicted peptide sequences.

Two additional technical changes were implemented in this version of pepStream. The first involves how the final docking score, used for ranking peptides, is obtained for each cMoRF. In the original pipeline, docking scores from all pairwise combinations of receptor and ligand conformations were accumulated into a list and averaged.

Notably, only some combinations of receptor and ligand conformation will produce stable complexes. [198] This was the reason for docking multiple conformations in the first place. Poor receptor/ligand combinations will yield low docking scores, whereas favorable combinations will yield high docking scores. Even if only one conformational combination is favorable, the peptide will still bind with the protein as long as both partners take the correct conformation *in vivo*. However, including these lower scores will lower the average docking score for the peptide, which may cause the peptide's ranking to be lower than it should be.

To address this, an alternative approach was devised. Instead of averaging all of the scores, the scores are first compiled into a list and sorted in descending order. Only the top $X\%$ of scores are averaged, representing the conformations with the strongest protein-peptide interactions. This method reduces bias from poorly-binding conformations.

Finally, modifications to the method for selecting the top-ranking peptides at the end of the pipeline was made. Previously, the docking scores for the target region were ranked and the top N peptides were considered as the best. However, this approach does not take into account specificity which is an important consideration

in pepStream. Specificity is difficult to quantify without experimental data, and so a surrogate a metric was introduced, given in Equation 5.1. This metric calculated the difference between the target docking score and the highest scoring alternate binding site. A positive value indicates that the peptide has specificity.

$$SP = SCORE_{target} - max(\{SCORE_{alt.sites}\}) \tag{5.1}$$

The target binding scores and specificity scores are ranked individually, resulting in two separate lists. Starting with 1 and adding a peptide each time, the intersection between the top N peptides from each list are found. When the length of the intersection reaches a user-defined length of candidate peptides, the algorithm finishes. This method balances target binding affinity with specificity within the final ranked list of peptides.

5.4 Results

PepStream was used to identify binding peptides for four MoRFs on beta-lactamase: the H3-H4 loop, the omega loop, the H9-H10 loop, and the H10-H11 loop. These loops were chosen as regions that exhibit functional dynamics in TEM beta-lactamase, as detailed in Chapter 4. That the functional motions correlate with enzyme substrate specificity suggests that binding a peptide to these locations may inhibit or shift these motions, allowing the catalytic properties of beta-lactamase to be controlled. These regions are shown on the structure of TEM-1 in Figure 5.2 in blue. For the analyses in this section, when one MoRF was chosen to be the target for pepStream, the remaining MoRFs were used as alternate targets for evaluating specificity.

Initial runs of pepStream over beta-lactamase using the original pipeline, detailed in Section 5.2, resulted in peptides that had some binding affinity for their target MoRF, but no selectivity. A lack of selectivity suggested that pepStream was not producing optimal peptides for beta-lactamase. Only the peptides designed for the

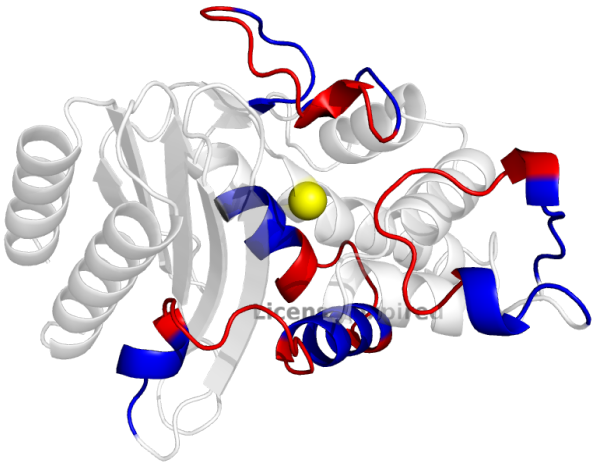

Figure 5.2: Structure of TEM-1 beta-lactamase colored by initial ten residue target MoRFs (red), and the expanded target MoRFs (blue). The catalytic SER70 is shown as a sphere in yellow.

H3-H4 loop showed specificity. However, peptides designed to bind to the other three MoRFs also had the highest affinity for the H3-H4 loop. This would imply that all peptides had specificity for the H3-H4 loop, regardless of what MoRF they were designed to bind to. Based on this, it could not be determined whether the peptides designed for the H3-H4 loop actually exhibited high specificity, or if this region had strong non-specific peptide-binding properties.

Although no valid peptides were generated from these runs, these results motivated many of the improvements that were made to the pipeline, as detailed in Section 5.3.1. After all the changes to the methodologies were implemented, pepStream was still unable to find specifically binding peptides for any MoRF other than the H3-H4 loop.

One potential explanation for why pepStream could not generate specific peptides was that the MoRF regions were too long. To test this, a new mode of running pepStream was formulated. Each MoRF was narrowed down to a 10 residue window. PepStream was run iteratively, extending the target each time by 1 residue on each side each iteration, until peptides with selectivity was found.

For each of the MoRFs used in this work, a "core ten" residue section was de-

termined by inspection and used as the starting window for the procedure. For the H3-H4, H9-H10, and H10-H11 loops, which represent short linkers between alpha helices, the core residues were selected to include the residues on the linkers. Additional residues were added on each side to make the total sequence length 10 in each case. For the omega loop, the core residues were chosen to include the part of the loop that overhangs the active site of the protein. The core residue sets for each MoRF are shown in red in Figure 5.2.

For each run of pepStream, a total of 9 alternate binding sites was used to compute specificity. To create alternate binding sites for this procedure, the full MoRF sequences were broken into roughly 10 residue windows. When a particular MoRF was chosen as the target, the alternate binding sites from that MoRF were retained, with the exception of the window that most overlapped with the actual target sequence. The actual active site of beta lactamase, centered on the Ser70 residue, with three flanking residues on either side, was also included in these 9 alternate sites. However, MEGADOCK failed in all cases to successfully bind any peptide with this region. This could be because Ser70 is relatively buried at the back of the binding pocket, making it difficult for a rigid docking scheme to accommodate a rigid peptide ligand structure in this region.

PepStream was run using the initial 10 residue fragment as a starting target. In an iterative process, a flanking residue on each side of the target was added and pepStream was re-run. This process was iterated until the target site was twenty residues long. By varying the target length, the effect of target length on the pepStream design process could be evaluated. When all iterations of pepStream had completed for each MoRF, the run which produced peptides with the highest specificity were selected to be further analyzed. No restrictions on the peptide length were enforced. However, the minimum fragment length was set to 7 because NCBI does not recommend blasting amino acid sequences with less than 7 residues.

Table 5.1: Number of unique cMoRF peptides found from searching the PDB for complimentary sequences

	H3-H4 Loop	Omega Loop	H9-H10 Loop	H10-H11 Loop
Length 10	13	4	1	1
Length 12	22	79	2	9
Length 14	55	221	30	77
Length 16	260	539	47	393
Length 18	908	1194	73	632
Length 20	1562	1631	151	1335

After all runs were completed, pepStream was able to construct peptides with affinity and specificity for all four MoRFs. The lengths resulting in the highest specificity peptides for each MoRF was 20 for the H3-H4 loop, 20 for the omega loop, 16 for the H9-H10 loop, and 20 for the H10-H11 loop. The number of cMoRF peptides found per-MoRF across all different lengths, for all four MoRFs, is shown in Table 5.1 and the affinity vs specificity graphs from the best runs are shown in Figure 5.3.

An interesting observation from Table 5.1 is that the H9-H10 loop generated substantially less cMoRFs compared to the other MoRFs. Interestingly, this is the only MoRF that does not directly border the active site. The number of cMoRFs found is dependent on the BLAST searches. One potential explanation for this observation is that there are fewer proteins in the PDB with homology for this region on beta-lactamase compared to the other MoRFs. With no obvious connection to the hydrolysis of antibiotics, there may be less evolutionary pressure to conserve the sequence of this loop between families of beta lactamases, resulting in the lower homology detected here.

In figure 5.3, special attention should be payed to the scales of both the x and y axes. The H3-H4 loop had significantly higher affinity and specificity compared to the other MoRFs. The other three MoRFs have roughly the same magnitude for affinity and for specificity compared to each other. In contrast to the prior pepStream results, peptides designed for other MoRFs did show specificity for the loop they were

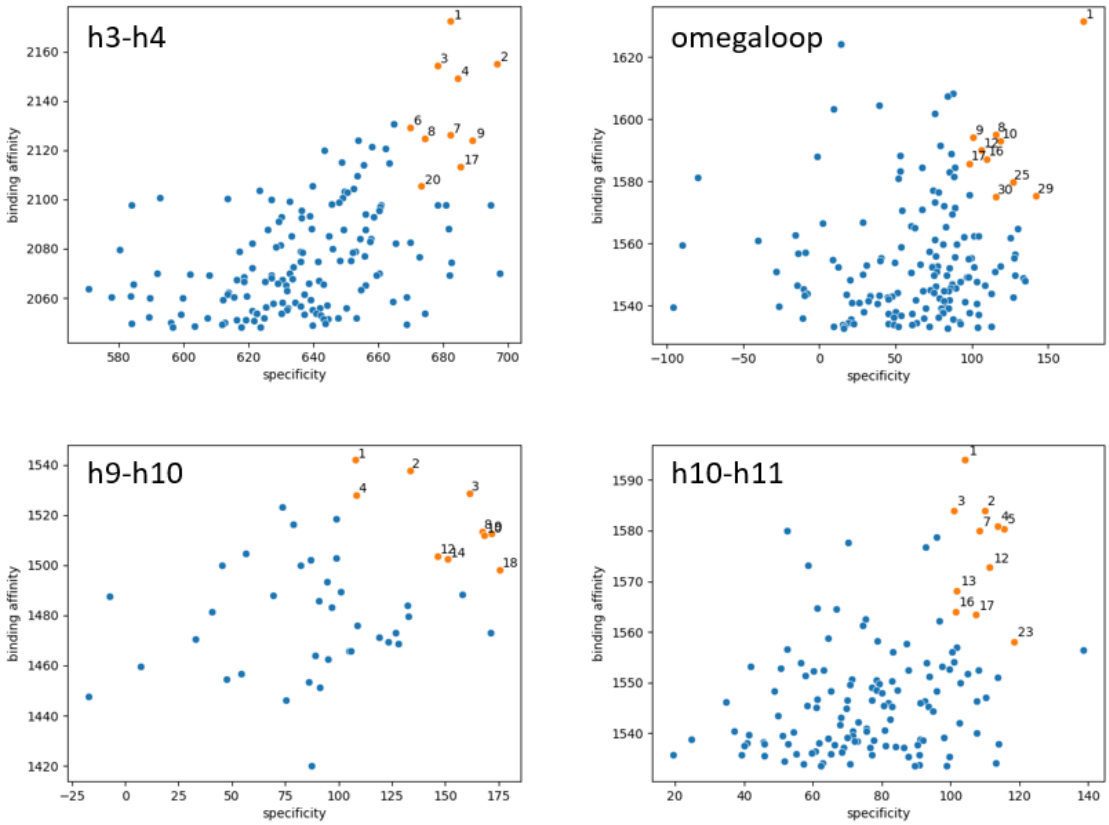

Figure 5.3: Affinity (y-axis) vs Specificity (x-axis) of peptides for a) the H3-H4 loop b) omega loop c) H9-H10 loop and d) H10-H11 loop. The yellow dots represent the top 10 peptides with overall best specificity and affinity for the target as determined by an automated algorithm. The labels represent their ranking in terms of binding affinity for the target.

designed to bind to. This suggests that the high affinity and specificity peptides show for the H3-H4 is significant, and that this loop has a strong affinity for binding peptides.

Notably, the H3-H4 loop is the site of the protein-protein interaction between TEM-1 and BLIP-I. The crystal structure of this complex (PDB: 3GMW [94]) shows that the interface between these proteins encompasses the entire H3-H4 loop, as well as a substantial portion of the TEM-1 molecular surface near the active site. An interesting study [108] attempted to try and derive a peptide from BLIP-I which would retain its binding and inhibitory properties for beta-lactamase. The fragment consisted of BLIP-I residues 46-51, which form a contact between an Alanine of BLIP-I and Tyr105 of the H3-H4 loop on TEM-1. Ultimately, this work showed that the peptide could inhibit beta-lactamase function, although at much lower potency compared to the full BLIP-I protein.

Breaking the alternate binding MoRFs into smaller windows enabled a more detailed look into the distribution of protein-protein interactions on beta-lactamase. Figure 5.4 presents the binding scores shown for each of the target and alternate target binding sites per MoRF. In each case, the binding scores for the target had a very narrow distribution in scores, likely because the scores in this list represent only the top N percent of docking scores.

The alternate target scores exhibit similar distributions for each pepStream run, despite being tested on different sets of peptides. For example, consider the three alternate MoRFs from the H3-H4 loop. In all four panels of Figure 5.4, the alternate targets for the H3-H4 loop have around the same value for binding affinity (around 1400). The same can be said for the other alternate MoRFs, which implies that the peptide binding properties on beta-lactamase are more strongly influenced by the protein sequence properties rather than the peptide sequence.

Examining the cMoRFs generated for each MoRF separately, no evident motifs or

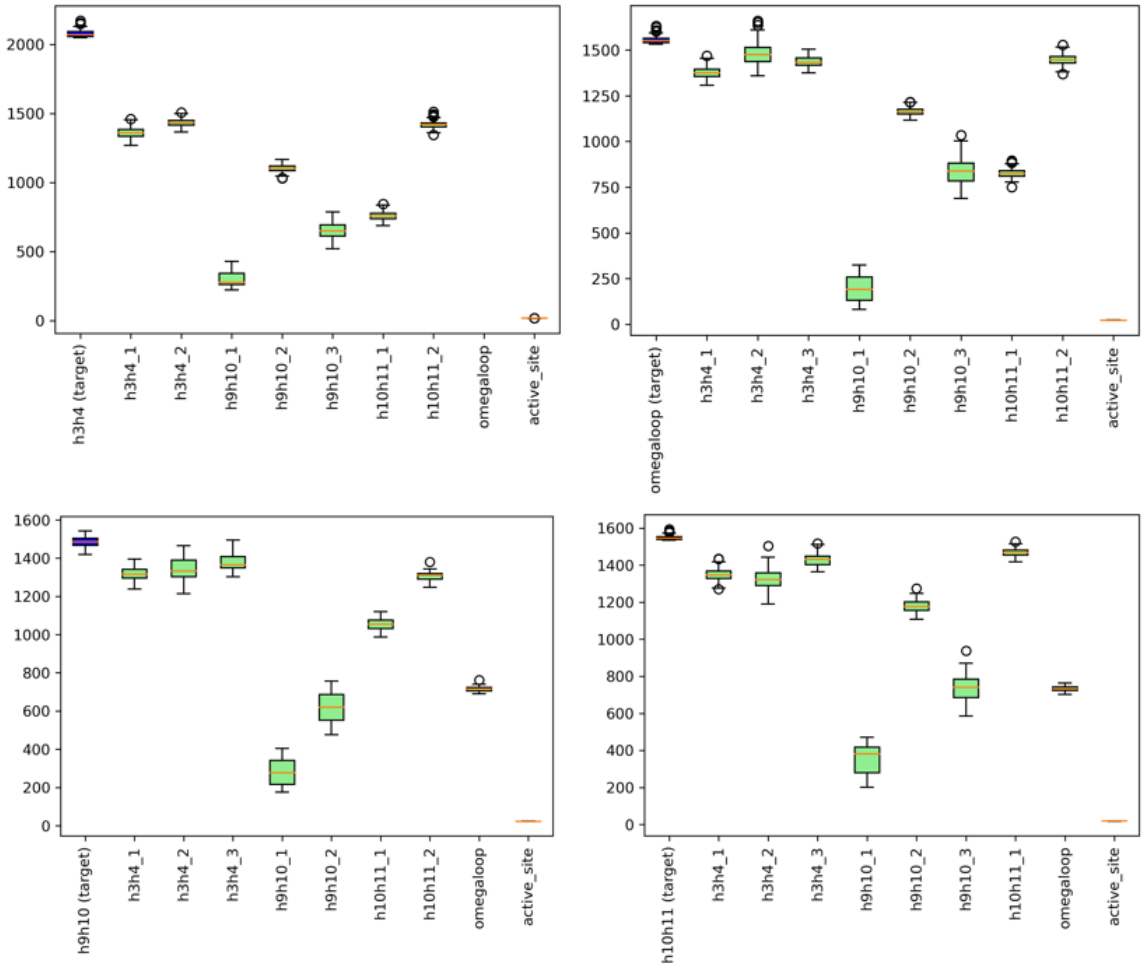

Figure 5.4: Comparing MEGADOCK scores for alternate-target binding sites used to compute specificity. For comparison, the target specificity scores are represented by the first boxplot in each panel.

sequence patterns that would indicate a general design principle were observed in the sequences generated by pepStream. Tracing the top cMoRFs back to the PDB entries they were extracted from can provide insights on the relationships between MoRF and cMoRF. In most cases the proteins from which the cMoRFs come from were unrelated to beta-lactamase. However, two of the top ten hits from the omega loop came from beta-lactamase relevant proteins. One of these hits came from a transcription regulator of the AmpC gene, which encodes AmpC beta-lactamase, and the other comes from ShyA an endopeptidase which is able to cleave the peptidoglycan cross-linking during the cell wall growth process which beta lactams inhibit. [199, 200] Additionally, one of the top ten hits for the H10-H11 loop came from a structure TEM-1 itself. [201]

The ProtParam module in BioPython [202] was used to compute six physical parameters for each of the top 10 and 100 peptides for each MoRF. The properties used were molecular weight, aromaticity, instability, hydrophobicity, isoelectric point (pI), and charge at pH 7.0. The results of this for the top 10 cMoRFs are shown in Figure 5.5.

These properties revealed that electrostatics (pI and charge) were primarily responsible for distinguishing peptides which bind with different MoRFs. The H3-H4 and omega loops prefer positively charged peptides, while the H9-H10 and H10-H11 loops prefer negatively charged peptides. Additionally, differences in hydrophobicity and molecular weight also observed between MoRFs. While there are some distinctions that can be made using only the top 10 peptides, most of these differences were not significant when the top 100 peptides (not shown) were considered.

5.4.1 cMoRF Sequence Analysis

Several methods were used to explore how pepStream constructed peptides to bind to the various MoRFs on beta-lactamase. Clustering was performed to explore differences between peptides designed for different MoRFs and similarities between peptides

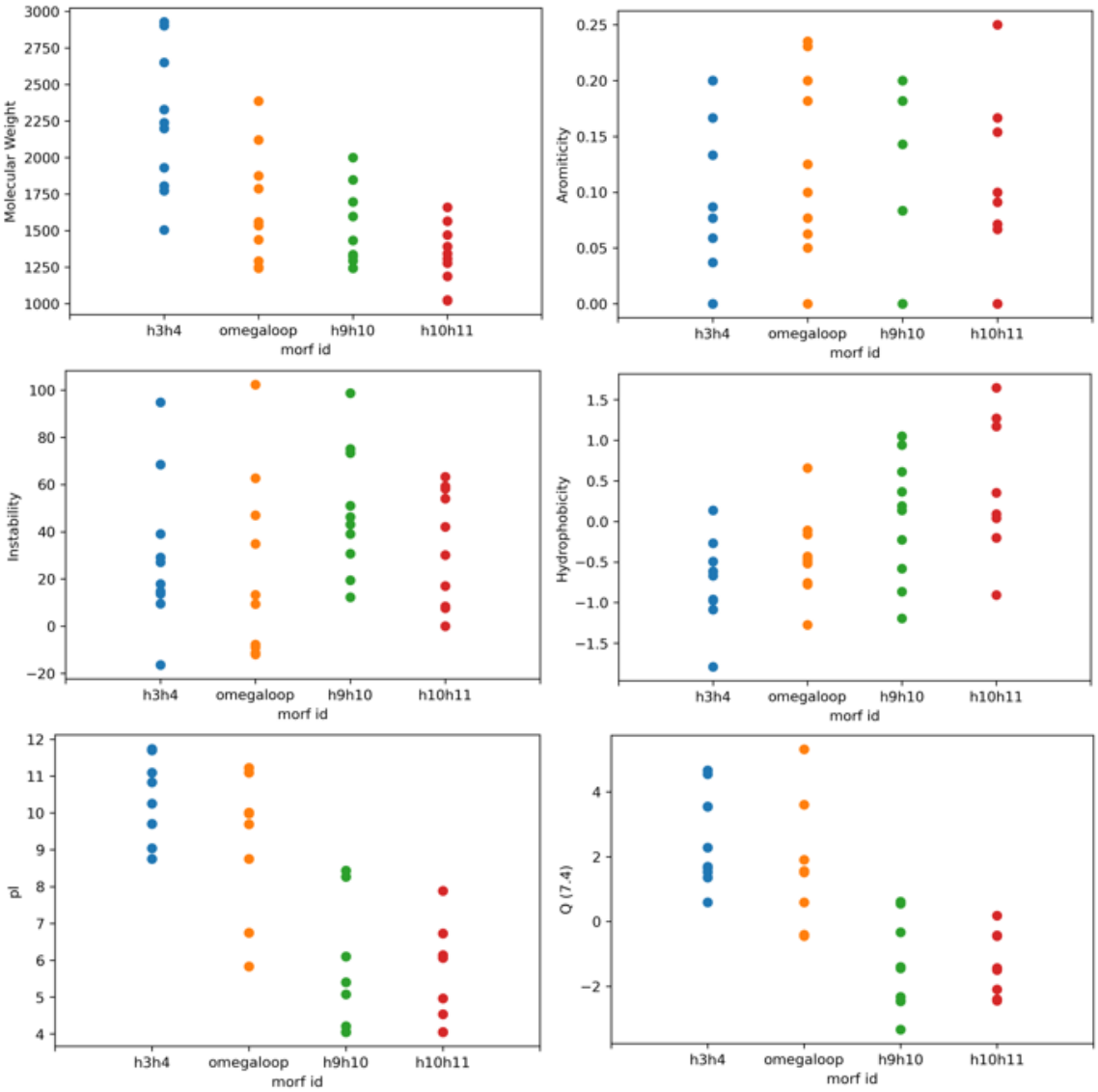

Figure 5.5: Peptide properties computed for the top 10 cMoRFs per MoRF.

designed for the same MoRF. Clustering peptide sequences between MoRFs aims to uncover general binding characteristics for each target region, while clustering peptide sequences designed for the same MoRF aimed to identify recurring motifs indicative of favorable binding characteristics.

The sequences were initially embedded into 480-dimensional vectors using the twelve layer evolutionary scale model, ESM-2, a protein language model. [203] ESM-2 is a transformer-based language model that has been trained using only sequence data, and these types of models have been shown to be able to be predictive of evolutionary and biophysical properties in proteins. [204] The embeddings were transformed using PCA and t-SNE (t-stochastic neighbor embedding [205]) to produce two dimensional maps to easily visualize the data.

To evaluate how well these embeddings differentiated peptide sequences designed for different MoRFs, clustering using k-means was performed directly on the embedded sequences and the two dimensional representations. The quality of clustering was evaluated using the silhouette score. A silhouette score of 1 indicates perfect clustering, a score of -1 indicates poor clustering, and a score of 0 indicates that the data samples are near the decision boundary between clusters. In addition to clustering based on protein language embeddings, a phylogenetic tree was constructed for the peptide sequences using the BLOSUM62 substitution matrix to explore whether the sequences could be distinguished through evolutionary clustering.

A maximum of the top 100 peptides (fewer, if less than 100 peptides were found) that exhibited specificity for each MoRF were combined into a single FASTA file and embedded using the ESM-2 model. The resulting embeddings were further transformed using PCA and t-SNE into low dimensional representations suitable for visualization. Only the top two components were kept on each model. For PCA, the top two modes were able to reconstruct around 70 percent of the original data variance.

Next, clustering using k-means was performed. These models used a k value of

4, for the 4 MoRFs where the peptides came from. When clustering directly on the embedded sequences the silhouette score was around 0.068. This indicates that the clustering was poor as samples were, on-average, close to the decision boundaries between clusters.

Alternatively, when clustering was performed on the top two components of the PCA and t-SNE transformed embeddings, the silhouette scores increased to 0.34 and 0.37 respectively which indicates the clustering improved. Visual inspection of the clusters does support this, as clusters predicted using the full ESM embeddings appear more overlapped (Figure 5.6 b,e) compared to clustering on the two dimensional slices (Figure 5.6 c,f).

Ultimately, the result is that neither clustering scheme correlates with the true labels (figure 5.6 a,d). This indicates that either the peptides are indistinguishable, despite being generated for different target MoRFs, or that the ESM-2 model is unable to capture the information underlying significant differences in the peptide sequences. One potential reason for this could be that a short peptide, such as the ones generated in pepStream, does not provide enough context for the transformer-based ESM-2 model to learn meaningful features.

The sequences were further tested using phylogenetic methods. Using the alignment software FAMSA [206], the top 10 peptide sequences for each MoRF were aligned and evolutionary distances were between sequences were computed using the BLOSUM62 substitution matrix. [207] A neighbor joining tree was constructed from these distances using BioPython and visualized using R (see Figure 5.7).

In figure 5.7, the resulting phylogenetic tree shows that the peptides were are not suitable for phylogenetic analysis. The MSA for this collection of sequences was poor, evidenced by the clade splits being bunched up near the root of the tree with long branch lengths to the leaves. This implies that if these peptides share any evolutionary relationship, they diverged from each other in the distant past. Despite being

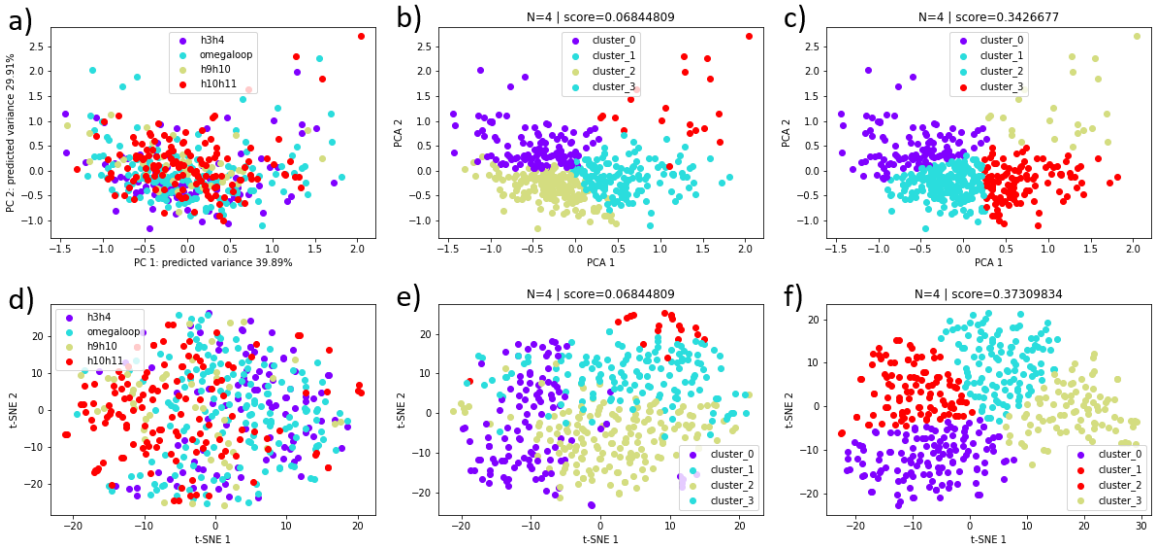

Figure 5.6: ESM encoding of peptide sequences. Scatter plots in a-c show the embedding vectors projected onto the top two PCA modes, while d-f show the vectors projected onto the top two t-SNE modes. Panels a and d are colored according to which MoRF each peptide was designed for, b and e are colored by clustering directly on the embedding vectors, while c and f are colored by clustering on the two dimensional representations.

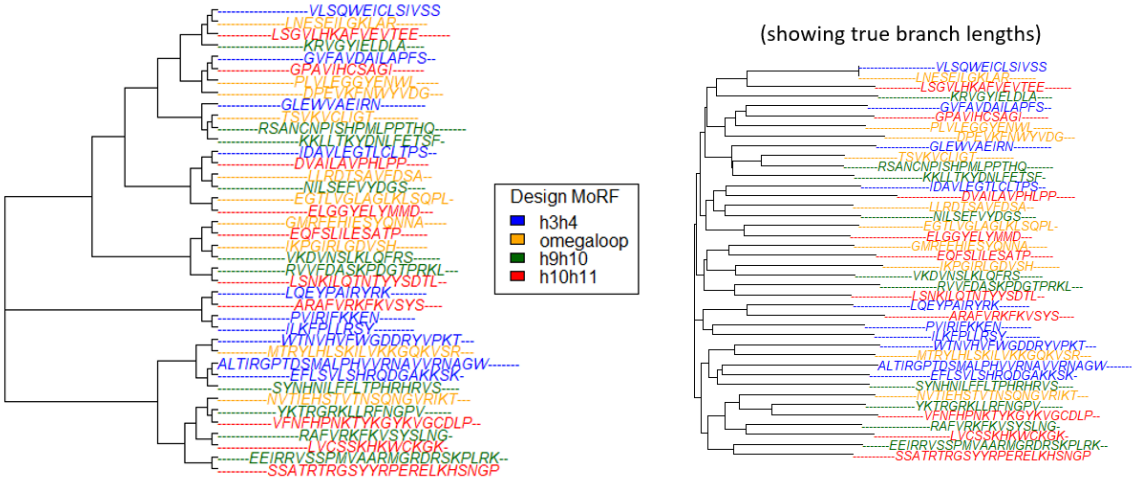

Figure 5.7: Phylogenetic tree for peptide sequences. Actual branch lengths are used to show the lack of evolutionary relationship between peptides designed for different MoRFs.

constructed from fragments of proteins known to interact with sequences homologous to the target MoRF, the peptides do not cluster by the MoRF they were designed for.

Despite being a null result, this suggests that that protein-protein interactions at these particular regions on beta-lactamase are not driven by evolution. In combination with the result from the ESM-2, this work did not not allow any patterns or motifs to be discerned for peptide binders to different regions in beta-lactamase. Although this result is not ideal, it informs the continued development of pepStream, highlighting the need to improve the pepStream sequence generation protocol. One possible approach would be to use a more targeted search for PDB structures where the amount of homology with the target region can be controlled more closely. Constructing cMoRFs from structures which are more relevant to the target protein may produce better candidates for binding.

5.5 Discussion and Future Outlook

The upgrades to pepStream have improved its ability to predict peptides with specificity to a particular MoRF on a protein. However, there still remains many improvements to be made in optimizing cMoRFs for protein-peptide interactions. The work described in this chapter represents an important step on the pathway to novel peptide inhibitors for beta-lactamase.

In this chapter a proof-of-concept methodology for designing binding peptides was demonstrated, and through extensive data analysis, several key areas in the pepStream methodology with the potential for future improvement were pinpointed. The two main areas include the cMoRF construction and diversification step and the docking step. It is important to recognize that pepStream was originally designed to predict binders for disordered proteins, which have different properties compared to globular proteins. Generalizing protocols to work for all proteins is a difficult challenge, and this is the first attempt to generalize the pepStream methodology outside

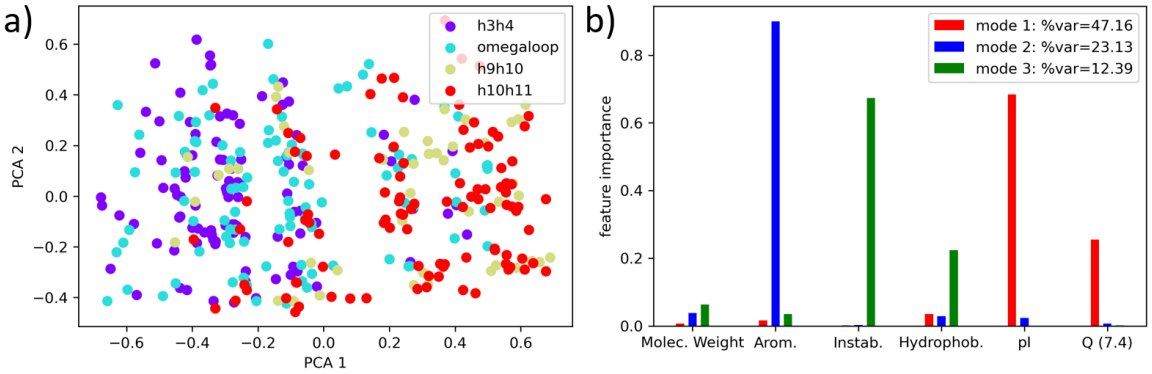

Figure 5.8: PCA results for physical properties of peptides. a) Top two PCA modes, as colored by the MoRF they were designed to bind to. b) Relative feature importance of each property toward each PCA mode. Importance is measured by the value of the squared loadings which correspond to the original property in the PC vector.

the scope of its intended purpose.

One of the notable results shown above was that peptides appeared to be indistinguishable from each other, regardless of the MoRF that peptides were designed for. Three clustering methods were attempted: by physical properties, by ESM-2 sequence embedding, and by phylogenetic tree clades. Among these methods, comparing the peptides by physical properties was the most effective.

The 6 physical properties of the top 100 peptides from each MoRF were visualized using PCA in Figure 5.8 a. The peptides generally exhibited a separation into two groups along PC1, which can be defined as either peptides designed for the H3-H4 and omega loops or peptides designed for the H9-H10 and H10-H11 loops. In agreement with the qualitative results above, the pI and charge at neutral pH most dominantly contributed to PC1, shown in Figure 5.8 b.

The rational design choice by pepStream becomes apparent when considering that the H3-H4 and omega loop fragments contain more charged/polar residues than the other two MoRFs, as can be seen in Figure 5.9. Ideally, a model using the physical properties of peptides could be constructed to uniquely differentiate the peptides designed for each of the four MoRFs. Global peptide properties, as calculated here, are unable to create such a model. Since the peptides exhibited specificity for their respec-

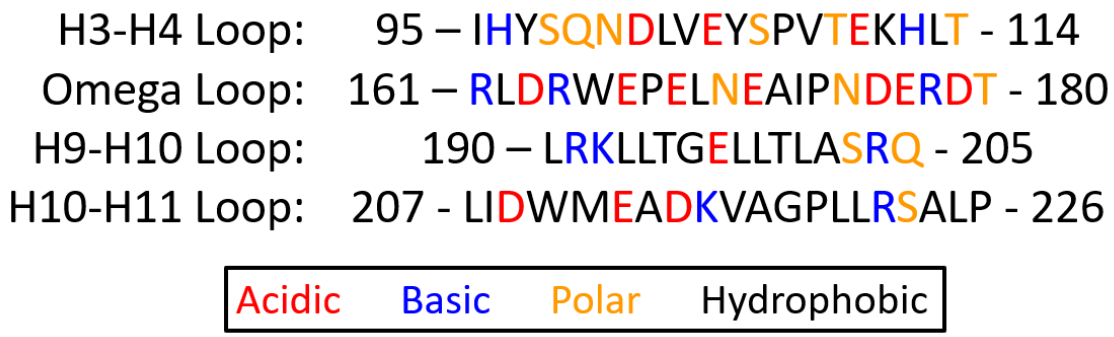

Figure 5.9: Sequences of the four MoRFs used in pepStream resulting in the highest specificity peptides. Amino acids are colored by biophysical properties.

tive MoRFs, it suggests that a set of features should exist that can describe the underlying differences between the peptides. Future studies will incorporate additional properties to investigate this, including local physical properties and environmental effects.

In this study, the peptide sequences generated by pepStream are not randomly generated, but are only weakly discernible as having been engineered to bind with a particular motif on a protein. One way in which this will be addressed in future work will focus on modifying the method by which seed sequences are used to generate the mutated footprints used to search the PDB.

Currently, unless a long target sequence is used, the footprints used in the the PSI-BLAST are typically very short, often approaching the 7 residue limit prescribed by NCBI. Given the complexity of protein sequence space, it is plausible that many proteins could contain subsequences with homology to a short fragment by random chance. Consequently, the PSI-BLAST will return many unrelated false-positive hits, biasing the later steps of pepStream. [208]

A proposed method for addressing this involves starting pepStream with an initial search of protein sequence space. This search uses a larger region around the target binding site to find a relatively large set of proteins with homology to the target. Using a larger sequence in this search will ensure that the returned proteins are not

random, however, have at least some, evolutionary relationship to the protein of interest. A user parameter could be introduced to control degree of diversity of this initial protein pool.

Next, all the proteins resulting from this search would be consolidated into a custom database to be used in the subsequent PSI-BLAST step. Everything after this step would run similar to the current implementation, including the BLASTp, where the full PDB would be searched. By limiting the PSI-BLAST to only search within proteins with sufficient homology to the target of interest, it is expected that the cMoRFs generated will have a higher likelihood to interact with the target.

Another avenue of improvement to pepStream pertains to the docking portion of the pipeline. While the changes already implemented as part of this work represent improvements to the prior scheme, ultimately using a rigid-rigid docking method is suboptimal for pepStream. Two major drawbacks of using MEGADOCK include: larger binding areas bias scores toward longer peptides and not considering flexibility when optimizing binding conformations.

An alternative method that was considered was AF2-mulitmer. [209] This model has emerged as one of the most accurate protein-protein interaction prediction models and has been widely adopted by the community as a docking method. This method predicts the receptor/ligand complex *de-novo*, therefore full flexibility is considered for the receptor and ligand. Additionally, AF2 has also found success in the realm of predicting peptide structure [210] and protein disordered regions [211]. Integrating AF2 into pepStream would allow the structure prediction and docking phases to be merged together as the docked complex could be directly predicted.

Ultimately, AF2-multimer has several drawbacks that make it also unsuitable for pepStream. Firstly, the prediction process is slow because of the MSA constructing phase, which is a major bottleneck of AF2 prediction time. Although there are some options to mitigate this, including using pre-computed MSAs or alternative

implementations such as colabfold [212], these options are not available on the UNCC HPC.

Secondly, while testing an implementation of pepStream that uses AF2, a strong bias was observed for predicting peptides to bind at the primary active site of the beta-lactamase. While this would be a positive result for protein-protein interactions in general, in pepStream, where synthetic peptides are designed to bind anywhere on the protein, including non-canonical binding sites on the receptor, this may cause the model to improperly predict where protein-peptide contacts are occurring. As further evidence of this, pLDDT, a measure of prediction confidence in AF2, for the TEM-1 chain was high, while for the peptide it was almost always low. This suggests that in most cases, AF2 did not know where to place the peptide relative to TEM-1.

Despite being essentially a black box model, AF2 relies on MSAs to predict a protein structure. [5] A neural network module called an evoformer transforms the MSA for each chain into a pair-wise distance constraint matrices, which are handed off to a structure module to predict the 3D structures. Co-evolution and conserved residues in the MSAs are used to predict where protein-protein contacts will occur and how the chains are oriented.

AF2 only uses sequence information as an input, and as a result, the model does not use physics information to inform how a polypeptide chain is structured. Instead, it learns how sequence information is encoded into an MSA and used learned correlations between residues to predict distance constraints. This approach leverages the massive protein databases that have been established over many years. [213, 214] While an MSA can be thought of as an indirect view of the biophysics associated with sequence dynamics, the capacity for AF2 to predict proteins structures is limited by the MSA. Notably, it has been observed that AF2 does not perform well on orphan proteins, which exhibit limited homology and produce shallow MSAs. [215]

These observations could be used to extrapolate an understanding of why AF2-

multimer is unable to predict correct binding poses for peptides generated in pepStream. Highly conserved residues are hallmarks of protein active sites, where a protein interacts with ligands, to ensure that the proteins function is conserved. [216] It is possible that AF2-multimer has rediscovered this rule through its training, and attempts to align the most conserved sites of each chain in 3D space.

In the case of two well-defined globular proteins with strong homologies and functional interactions, this assumption is valid. However, if a protein with strong homology and a synthetic peptide with a little to no homology were to be predicted, as is the case with beta-lactamase and peptides generated by pepStream, AF2-multimer would only be able to find conserved residues for the protein. By its own rules, the model would try and place the peptides near these conserved protein residues. This would explain what was observed when AF2-multimer was tested on pepStream datasets.

While this drawback exists in the current model, it could be overcome in future versions of the model by incorporating a more curated training set for proteins and peptides with shallow MSAs, and optimizing hardware to speed up predictions. For these reasons, AF2-multimer is an attractive, yet as of now unsuitable, alternative for pepStream.

Finally, a potential improvement for mitigating the sequence length dependence of MEGADOCK scores is described here. Figure 5.10 demonstrates the dependence of the MEGADOCK scoring function on sequence length for global docking. In the updated implementation of pepStream, the solution was to reorganize the docking phase of the pipeline to remove global docking altogether. Despite this, a low level of correlation can still exist in the restrained docking scores. As a proposed future improvement to pepStream, this correlation can be corrected by applying the proper transformation to the raw docking scores.

Several methods were explored to normalize MEGADOCK scores so as to minimize the correlation between MEGADOCK score and peptide sequence length. Initially, a

simple attempt was to directly normalize the scores by dividing by the length of the peptide sequence, shown in Equation 5.2.

$$SCORE_{norm}(peptide) = \frac{SCORE_{raw}(peptide)}{length(peptide)} \tag{5.2}$$

This approach resulted in a strong negative correlation between MEGADOCK scores and sequence length, suggesting that a weaker adjustment was required. An empirical approach was designed to find the optimal value for a parameter N, by sweeping over a range of values to find to one that minimizes correlation between scores and sequence length. The resulting normalisation scheme is given by Equation 5.3.

$$SCORE_{norm}(peptide) = \frac{SCORE_{raw}(peptide)}{length(peptide)^{(1/N)}} \tag{5.3}$$

N is allowed to go from 0.0 to 100, excluding 0.0 for stability, to test powers ranging from infinity to 0.01. Figure 5.10 shows the correlation curve as a function of N, as well as the MEGADOCK scores before and after transformation. The longest peptides were the strongest affected. Before permanent implementation into pepStream, it will be important to benchmark this parameter N across several datasets to understand how normalization impacts the overall predictions of pepStream, or to see if any heuristic exists for choosing the value of N.

5.6 Conclusion

In this chapter a significant update to the practical implementation and methodology of the peptide design pipeline, pepStream, is described. These improvements include consolidating the pipeline into a single SLURM script with a user-friendly input parameter file interface, incorporating modern structure prediction and docking programs to increase the speed of the pipeline, devising a new scheme for running pepStream, and various other minor technical improvements.

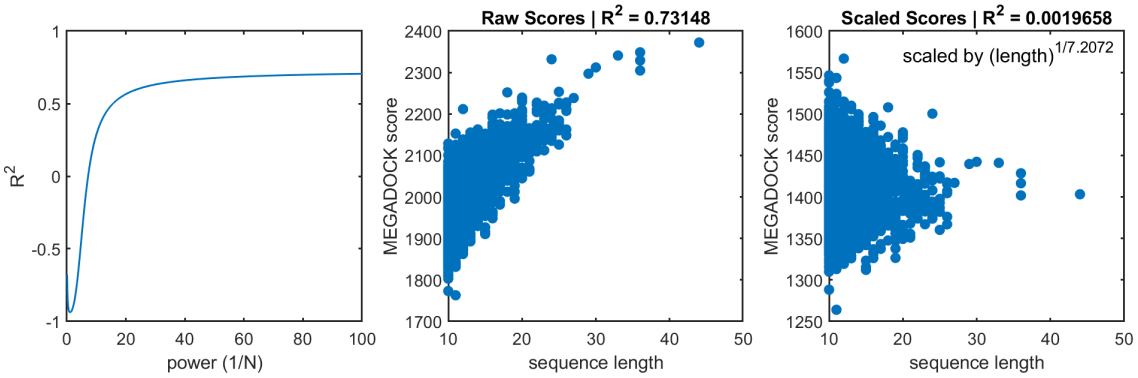

Figure 5.10: Docking scores vs sequence length from MEGADOCK before and after transformation by the optimal N value to reduce correlations. a) correlation as a function of the N value. b) docking scores before transformation. c) docking scores after optimal transformation.

The latest version of pepStream was applied to predicting novel peptide binders for beta-lactamase with the goal of finding a new generation of peptide-based beta-lactamase inhibitors, which operate by targeting regions on the enzyme which exhibit functional dynamics required for substrate recognition. The results show that pepStream was able to identify peptides that would bind at various region with specificity, however the specific properties that underlie their binding were unable to be determined. These results serve as an important benchmark for understanding how pepStream can be further improved in future work.

Among the four regions targeted in this study, which exhibit functional motions, the H3-H4 loop showed notably high affinity for protein-peptide interactions. This suggests that the H3-H4 loop might be a promising potential target for the future design of peptide beta-lactamase inhibitors.

The results presented contribute to understanding how pepStream performs in practice, especially on well ordered globular proteins. From this, two directions of future work for pepStream have been identified: improving the sequence search in order to generate higher quality peptides and improving the docking procedures to incorporate flexibility and unbiased scoring metrics for protein-peptide complexes.

This work will surely lead to a faster and more accurate pepStream, which will be a great aid in the search for novel peptide drugs.

CHAPTER 6: CONCLUSIONS

Throughout this dissertation, the function of beta-lactamase has been probed through the lens of protein dynamics. Here, beta-lactamase function refers to the differences in catalytic efficiency of different enzymes for various antibiotics. Just as these functional differences are subtle between enzymes, the mechanisms controlling them are also subtle, and elucidating such changes represents an important challenge in biophysics and bioinformatics.

Experiments can provide much information about differences in beta-lactamase efficiency toward different antibiotic drugs. However, this information often demonstrates that differences exist, but does not provide mechanistic details about how these changes in function arise. Computational approaches can be leveraged to better understand the details of how protein function occurs. Molecular Dynamics simulations provide atomistic details about molecular systems that can be used to probe the detailed motions of proteins. Through MD, phenomena in protein systems like conformational change, dynamic allostery, and molecular docking can be directly observed.

Many methods designed for interpreting dynamics in MD simulations focus on the largest motions of the system, as previously it has been assumed that these were the most important. However, function can also be controlled by small-scale changes to protein conformations, through mechanisms that can be difficult to elucidate from the background noise of the simulation. As part of this dissertation, novel approaches to identifying the functional dynamics in proteins were developed and tested on beta-lactamase to elucidate the dynamic changes in the enzyme that control substrate recognition and binding.

Changes in the motion of beta-lactamase were observed to correlate to its substrate recognition properties. It is possible that modulating these motions can also modulate the binding specificity of the enzyme. This suggests the possibility of a novel form of beta-lactamase inhibition, where inhibitors target regions of the enzyme that couple to functional dynamics, rather than directly targeting the catalytic Ser70.

Methods of designing peptides which are capable of binding specifically with these regions were tested. The approach used a pipeline called pepStream, which was designed to predict peptide binders for disordered regions on proteins. In the work presented here, several modifications to pepStream were proposed and implemented, and peptide binders for four different MoRFs on beta-lactamase were predicted.

6.1 Summary Of Results

Functional dynamics in beta-lactamase exist in the form of small-scale motion and local conformational changes. Mutations to the enzyme imparts flexibility into loops which border the active site, which in turn allows the enzyme to accommodate different types of ligands in the pocket. These changes were expressed as unique dynamic signatures induced in beta-lactamase enzymes according to the specific combination of mutations and antibiotic ligand present.

In particular, changes in motions to the omega and H10-H11 loop were observed to be correlated with changing substrate specificity. The H10-H11 loop was particularly dynamic, which suggested that the increased flexibility allowed it to adopt different conformations, which in turn allows the loop to help stabilize antibiotic molecules during hydrolysis.

Another loop between helix 9 and helix 10 was also noted to exhibit dynamic changes resulting from both mutation and antibiotic binding. Previously, this loop was not known to directly impact the catalytic function of the enzyme, as it is on the opposite side of the protein from the active site. Based on these observations, the H9-H10 loop is proposed to be involved in a general mechanism aimed at maintaining

the stability of beta-lactamase when changes are induced elsewhere on the enzyme, which directly change substrate specificity.

The long range effects of changes in motion impacting the binding affinity of beta-lactamase for different antibiotics is an example of dynamic allostery. Allostery is a powerful mechanism which exploits the cooperativity of motions in proteins. Using a method for identifying dynamic allostery signals in proteins, hot spots of dynamic allostery were identified on beta-lactamase. In addition to verifying a known allosteric signal around helix 11, the H3-H4 loop was also found to have a strong allosteric signal coupling a rigidification of motion in this loop with binding efficacy in the active site.

The various dynamic signatures in beta-lactamase identified through this work demonstrate the importance of protein dynamics in controlling beta-lactamase activity. To test whether antibiotic resistance could be identified solely using these motions, the likelihood that enzymes would express an expanded substrate recognition for extended spectrum antibiotics was predicted. This likelihood was found using a classifier trained on MD simulations of proteins which do and do not bind well to third generation cephalosporins. It was found that the classifier could correctly rank enzymes in terms of likelihood not to bind based on apo simulations alone.

Finally a proof of concept for designing peptides to inhibit beta-lactamase at regions correlated to functional dynamics was demonstrated. Using a modified pepStream, peptides which bound to the H3-H4, omega, H9-H10, and H10-H11 loops were generated. These peptides showed both high affinity and specificity for binding with the specific region of beta-lactamase they were designed to bind with. Although the properties that imparted the peptides with specificity for different regions of beta-lactamase could not be identified, this first step in the design of a new class of beta-lactamase inhibitors represents a step forward for both combating antibiotic resistance in beta-lactamase and the rational design of peptide drugs.

6.2 Discussion and Impact

The nature of functional dynamics in beta-lactamase would not have been able to be discerned without the methodological advancement of SPLOC. Unlike other MD analysis techniques, SPLOC does not rely on assumptions to elucidate changes in motions, rather it uses a supervised comparison to learn a data-driven hypothesis, which is described by discriminating features called d-modes. By comparing simulations using emergent properties of data packets, rather than a conformation-by-conformation comparison approach, SPLOC is able to identify statistically significant changes in motion between enzymes, even when the two enzymes share the majority of their conformational spaces.

The dynamic signatures underlying substrate recognition reveals the complexity of beta-lactamase function. Enzyme specificity is dictated by its specific mutation combination through the introduction of flexibility around the binding pocket. When the right antibiotic is present, this flexibility is used to adjust the enzymes conformation in local regions to optimize the interaction.

The ability to classify antibiotic resistance phenotype based on the motions sampled in an apo enzyme is remarkable, and highlights the potential applications of functional dynamics. The classification scheme here is currently intractable for real-time antibiotic resistance detection, however, the capacity for MD is becoming more accessible due to better hardware and machine learning-based forcefields. In the future, a scheme similar to what was presented here could be implemented on a larger scale to predict antibiotic resistance to commonly used beta lactam drugs. In this scheme the classifiers can be precomputed using SPLOC, based on simulations of known enzyme/drug pairs, and only the novel beta-lactamase would need to be simulated to make a prediction. Testing enzymes in this way can reduce the number of experiments needed to determine what antibiotics will be needed to effectively treat a bacterial infection.

The different approaches used in this work were performed on beta-lactamase systems independently, and the information gleaned from one method was used only as an insight for another. Although it is beyond the scope of this work, an integrated approach in which methods like SPLOC and pepStream are able to directly interact and pass information to each other is feasible based on what was shown here. In this grand computational drug discovery pipeline, drugs for inhibiting beta-lactamase could be generated by pepStream, and optimized using SPLOC or the allostery program to induce or inhibit specific motions in the enzymes.

This integrated approach represents a new paradigm for fighting antibiotic resistance. The computations tools described in this work form the basis of this paradigm, where the connection between protein dynamics and protein function can be fully exploited to produce novel drugs.

REFERENCES

[1] C. Avery, L. Baker, and D. J. Jacobs, "Functional dynamics of substrate recognition in tem beta-lactamase," *Entropy*, vol. 24, no. 5, p. 729, 2022.

[2] T. Grear, C. Avery, J. Patterson, and D. J. Jacobs, "Molecular function recognition by supervised projection pursuit machine learning," *Scientific reports*, vol. 11, no. 1, p. 4247, 2021.

[3] B. Alberts, *Molecular biology of the cell*. Garland science, 2017.

[4] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973.

[5] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[6] B. Rost, "Twilight zone of protein sequence alignments," *Protein Eng*, vol. 12, no. 2, pp. 85–94, 1999.

[7] C. A. Orengo, D. T. Jones, and J. M. Thornton, "Protein superfamilles and domain superfolds," *Nature*, vol. 372, no. 6507, pp. 631–634, 1994.

[8] P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, *et al.*, "A large-scale evaluation of computational protein function prediction," *Nature methods*, vol. 10, no. 3, pp. 221–227, 2013.

[9] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[10] K. Henzler-Wildman and D. Kern, "Dynamic personalities of proteins," *Nature*, vol. 450, no. 7172, pp. 964–972, 2007.

[11] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, "The energy landscapes and motions of proteins," *Science*, vol. 254, no. 5038, pp. 1598–1603, 1991.

[12] C. J. Oldfield and A. K. Dunker, "Intrinsically disordered proteins and intrinsically disordered protein regions," *Annual review of biochemistry*, vol. 83, pp. 553–584, 2014.

[13] D. Chin and A. R. Means, "Calmodulin: a prototypical calcium sensor," *Trends in cell biology*, vol. 10, no. 8, pp. 322–328, 2000.

[14] H. N. Motlagh, J. O. Wrabl, J. Li, and V. J. Hilser, "The ensemble nature of allostery," *Nature*, vol. 508, no. 7496, pp. 331–339, 2014.

[15] CDC, "Antibiotic resistance threats in the united states, 2019," 2019.

[16] I. Sultan, S. Rahman, A. T. Jan, M. T. Siddiqui, A. H. Mondal, and Q. M. R. Haq, "Antibiotics, resistome and resistance mechanisms: A bacterial perspective," *Frontiers in microbiology*, vol. 9, p. 2066, 2018.

[17] F. C. Tenover, "Development and spread of bacterial resistance to antimicrobial agents: an overview," *Clinical infectious diseases*, vol. 33, no. Supplement_3, pp. S108–S115, 2001.

[18] L. Rizzo, C. Manaia, C. Merlin, T. Schwartz, C. Dagot, M. Ploy, I. Michael, and D. Fatta-Kassinos, "Urban wastewater treatment plants as hotspots for antibiotic resistant bacteria and genes spread into the environment: a review," *Science of the total environment*, vol. 447, pp. 345–360, 2013.

[19] S. Doron and L. E. Davidson, "Antimicrobial stewardship," in *Mayo Clinic Proceedings*, vol. 86, pp. 1113–1123, Elsevier, 2011.

[20] V. Blay, B. Tolani, S. P. Ho, and M. R. Arkin, "High-throughput screening: todayâs biochemical and cell-based approaches," *Drug Discovery Today*, vol. 25, no. 10, pp. 1807–1821, 2020.

[21] R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, *et al.*, "Impact of high-throughput screening in biomedical research," *Nature reviews Drug discovery*, vol. 10, no. 3, pp. 188–195, 2011.

[22] K. Bush, "Past and present perspectives on β-lactamases," *Antimicrobial agents and chemotherapy*, vol. 62, no. 10, pp. 10–1128, 2018.

[23] K. Bush and P. A. Bradford, "β-lactams and β-lactamase inhibitors: an overview," *Cold Spring Harbor perspectives in medicine*, vol. 6, no. 8, 2016.

[24] I. Massova and S. Mobashery, "Kinship and diversification of bacterial penicillin-binding proteins and β-lactamases," *Antimicrobial agents and chemotherapy*, vol. 42, no. 1, pp. 1–17, 1998.

[25] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[26] P. Karplus and G. Schulz, "Prediction of chain flexibility in proteins: a tool for the selection of peptide antigens," *Naturwissenschaften*, vol. 72, no. 4, pp. 212–213, 1985.

[27] G. W. Tumbic, M. Y. Hossan, and M. C. Thielges, "Protein dynamics by two-dimensional infrared spectroscopy," *Annual Review of Analytical Chemistry*, vol. 14, pp. 299–321, 2021.

[28] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, "Systematic validation of protein force fields against experimental data," *PloS one*, vol. 7, no. 2, p. e32131, 2012.

[29] P. C. Souza, R. Alessandri, J. Barnoud, S. Thallmair, I. Faustino, F. Grünewald, I. Patmanidis, H. Abdizadeh, B. M. Bruininks, T. A. Wassenaar, *et al.*, "Martini 3: a general purpose force field for coarse-grained molecular dynamics," *Nature methods*, vol. 18, no. 4, pp. 382–388, 2021.

[30] Y. I. Yang, Q. Shao, J. Zhang, L. Yang, and Y. Q. Gao, "Enhanced sampling in molecular dynamics," *The Journal of chemical physics*, vol. 151, no. 7, 2019.

[31] G. Sormani, A. Rodriguez, and A. Laio, "Explicit characterization of the free-energy landscape of a protein in the space of all its cα carbons," *Journal of chemical theory and computation*, vol. 16, no. 1, pp. 80–87, 2019.

[32] S. W. Englander and L. Mayne, "The nature of protein folding pathways," *Proceedings of the National Academy of Sciences*, vol. 111, no. 45, pp. 15873–15880, 2014.

[33] K. A. Henzler-Wildman, M. Lei, V. Thai, S. J. Kerns, M. Karplus, and D. Kern, "A hierarchy of timescales in protein dynamics is linked to enzyme catalysis," *Nature*, vol. 450, no. 7171, pp. 913–916, 2007.

[34] C. C. David and D. J. Jacobs, "Principal component analysis: a method for determining the essential dynamics of proteins," *Protein dynamics: Methods and protocols*, pp. 193–226, 2014.

[35] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 32, no. 5, pp. 922–923, 1976.

[36] C. Avery, J. Patterson, T. Grear, T. Frater, and D. J. Jacobs, "Protein function analysis through machine learning," *Biomolecules*, vol. 12, no. 9, p. 1246, 2022.

[37] A. R. Atilgan, S. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, "Anisotropy of fluctuation dynamics of proteins with an elastic network model," *Biophysical journal*, vol. 80, no. 1, pp. 505–515, 2001.

[38] V. Tozzini, "Coarse-grained models for proteins," *Current opinion in structural biology*, vol. 15, no. 2, pp. 144–150, 2005.

[39] I. Bahar, A. R. Atilgan, and B. Erman, "Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential," *Folding and Design*, vol. 2, no. 3, pp. 173–181, 1997.

[40] J. R. Taylor and J. R. Taylor, *Classical mechanics*, vol. 1. Springer, 2005.

[41] A. S. Ettayapuram Ramaprasad, S. Uddin, J. Casas-Finet, and D. J. Jacobs, "Decomposing dynamical couplings in mutated scfv antibody fragments into stabilizing and destabilizing effects," *Journal of the American Chemical Society*, vol. 139, no. 48, pp. 17508–17517, 2017.

[42] A. Amadei, A. B. Linssen, and H. J. Berendsen, "Essential dynamics of proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 17, no. 4, pp. 412–425, 1993.

[43] M. Rueda, P. Chacón, and M. Orozco, "Thorough validation of protein normal mode analysis: a comparative study with essential dynamics," *Structure*, vol. 15, no. 5, pp. 565–575, 2007.

[44] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

[45] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[46] A. Kitao and N. Go, "Investigating protein dynamics in collective coordinate space," *Current opinion in structural biology*, vol. 9, no. 2, pp. 164–169, 1999.

[47] G. Fiorin, M. L. Klein, and J. Hénin, "Using collective variables to drive molecular dynamics simulations," *Molecular Physics*, vol. 111, no. 22-23, pp. 3345–3362, 2013.

[48] R. B. Cattell and S. Vogelmann, "A comprehensive trial of the scree and kg criteria for determining the number of factors," *Multivariate Behavioral Research*, vol. 12, no. 3, pp. 289–325, 1977.

[49] C. C. David, E. R. A. Singam, and D. J. Jacobs, "Jed: a java essential dynamics program for comparative analysis of protein trajectories," *BMC bioinformatics*, vol. 18, pp. 1–9, 2017.

[50] P. J. Huber, "Projection pursuit," *The annals of Statistics*, pp. 435–475, 1985.

[51] S. Nanga, A. T. Bawah, B. A. Acquaye, M.-I. Billa, F. D. Baeta, N. A. Odai, S. K. Obeng, and A. D. Nsiah, "Review of dimension reduction methods," *Journal of Data Analysis and Information Processing*, vol. 9, no. 3, pp. 189–231, 2021.

[52] F. Sittel, A. Jain, and G. Stock, "Principal component analysis of molecular dynamics: On the use of cartesian vs. internal coordinates," *The Journal of Chemical Physics*, vol. 141, no. 1, 2014.

[53] D. M. Van Aalten, B. L. De Groot, J. B. Findlay, H. J. Berendsen, and A. Amadei, "A comparison of techniques for calculating protein essential dynamics," *Journal of Computational Chemistry*, vol. 18, no. 2, pp. 169–181, 1997.

[54] C. C. David, C. S. Avery, and D. J. Jacobs, "Jedi: Java essential dynamics inspectorâa molecular trajectory analysis toolkit," *BMC bioinformatics*, vol. 22, no. 1, p. 226, 2021.

[55] J. Liu and R. Nussinov, "Allostery: an overview of its history, concepts, methods, and applications," *PLoS computational biology*, vol. 12, no. 6, p. e1004966, 2016.

[56] A. Del Sol, C.-J. Tsai, B. Ma, and R. Nussinov, "The origin of allosteric functional modulation: multiple pre-existing pathways," *Structure*, vol. 17, no. 8, pp. 1042–1050, 2009.

[57] K. Gunasekaran, B. Ma, and R. Nussinov, "Is allostery an intrinsic property of all dynamic proteins?," *Proteins: Structure, Function, and Bioinformatics*, vol. 57, no. 3, pp. 433–443, 2004.

[58] M. H. Ahmed, M. S. Ghatge, and M. K. Safo, "Hemoglobin: structure, function and allostery," *Vertebrate and invertebrate respiratory proteins, lipoproteins and other body fluid proteins*, pp. 345–382, 2020.

[59] D. D. Boehr, R. Nussinov, and P. E. Wright, "The role of dynamic conformational ensembles in biomolecular recognition," *Nature chemical biology*, vol. 5, no. 11, pp. 789–796, 2009.

[60] J. Monod, J. Wyman, and J.-P. Changeux, "On the nature of allosteric transitions: a plausible model," *J Mol Biol*, vol. 12, no. 1, pp. 88–118, 1965.

[61] D. E. Koshland Jr, G. Némethy, and D. Filmer, "Comparison of experimental binding data and theoretical models in proteins containing subunits," *Biochemistry*, vol. 5, no. 1, pp. 365–385, 1966.

[62] G. Weber, "Ligand binding and internal equilibiums in proteins," *Biochemistry*, vol. 11, no. 5, pp. 864–878, 1972.

[63] A. Cooper and D. Dryden, "Allostery without conformational change: a plausible model," *European Biophysics Journal*, vol. 11, pp. 103–109, 1984.

[64] CDC, "Outpatient antibiotic prescriptions - united states, 2020," 2020.

[65] A. Fleming, "On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of b. influenzae," *British journal of experimental pathology*, vol. 10, no. 3, p. 226, 1929.

[66] E. Chain, H. W. Florey, A. D. Gardner, N. G. Heatley, M. A. Jennings, J. Orr-Ewing, and A. G. Sanders, "Penicillin as a chemotherapeutic agent," *The lancet*, vol. 236, no. 6104, pp. 226–228, 1940.

[67] E. Abraham and P. Loder, "Cephalosporin c," *Cephalosporins and penicillins: chemistry and biology*, pp. 2–26, 1972.

[68] P. M. Blumberg and J. L. Strominger, "Interaction of penicillin with the bacterial cell: penicillin-binding proteins and penicillin-sensitive enzymes," *Bacteriological reviews*, vol. 38, no. 3, pp. 291–335, 1974.

[69] D. J. Waxman and J. L. Strominger, "Penicillin-binding proteins and the mechanism of action of beta-lactam antibiotics," *Annual review of biochemistry*, vol. 52, no. 1, pp. 825–869, 1983.

[70] M. S. Wilke, A. L. Lovering, and N. C. Strynadka, "β-lactam antibiotic resistance: a current structural perspective," *Current opinion in microbiology*, vol. 8, no. 5, pp. 525–533, 2005.

[71] R. P. Ambler, "The structure of β-lactamases," *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, vol. 289, no. 1036, pp. 321–331, 1980.

[72] K. Bush, G. A. Jacoby, and A. A. Medeiros, "A functional classification scheme for beta-lactamases and its correlation with molecular structure," *Antimicrobial agents and chemotherapy*, vol. 39, no. 6, pp. 1211–1233, 1995.

[73] K. Bush and G. A. Jacoby, "Updated functional classification of β-lactamases," *Antimicrobial agents and chemotherapy*, vol. 54, no. 3, pp. 969–976, 2010.

[74] T. Palzkill, "Metallo-β-lactamase structure and function," *Annals of the New York Academy of Sciences*, vol. 1277, no. 1, pp. 91–104, 2013.

[75] H. Christensen, M. T. Martin, and S. G. Waley, "Beta-lactamases as fully efficient enzymes. determination of all the rate constants in the acyl-enzyme mechanism.," *Biochemical Journal*, vol. 266, no. 3, p. 853, 1990.

[76] A. Egorov, M. Rubtsova, V. Grigorenko, I. Uporov, and A. Veselovsky, "The role of the ω-loop in regulation of the catalytic activity of tem-type β-lactamases," *Biomolecules*, vol. 9, no. 12, p. 854, 2019.

[77] P.-Y. Savard and S. M. Gagné, "Backbone dynamics of tem-1 determined by nmr: evidence for a highly ordered protein," *Biochemistry*, vol. 45, no. 38, pp. 11414–11424, 2006.

[78] S. Morin and S. M. Gagné, "Nmr dynamics of pse-4 β-lactamase: an interplay of ps-ns order and μs-ms motions in the active site," *Biophysical journal*, vol. 96, no. 11, pp. 4681–4691, 2009.

[79] J. R. Porter, K. E. Moeder, C. A. Sibbald, M. I. Zimmerman, K. M. Hart, M. J. Greenberg, and G. R. Bowman, "Cooperative changes in solvent exposure identify cryptic pockets, switches, and allosteric coupling," *Biophysical Journal*, vol. 116, no. 5, pp. 818–830, 2019.

[80] S. Banerjee, U. Pieper, G. Kapadia, L. K. Pannell, and O. Herzberg, "Role of the ω-loop in the activity, substrate specificity, and structure of class a β-lactamase," *Biochemistry*, vol. 37, no. 10, pp. 3286–3296, 1998.

[81] C. L. Tooke, P. Hinchliffe, E. C. Bragginton, C. K. Colenso, V. H. Hirvonen, Y. Takebayashi, and J. Spencer, "β-lactamases and β-lactamase inhibitors in the 21st century," *Journal of molecular biology*, vol. 431, no. 18, pp. 3472–3500, 2019.

[82] D. Sirot, J. Sirot, R. Labia, A. Morand, P. Courvalin, A. Darfeuille-Michaud, R. Perroux, and R. Cluzel, "Transferable resistance to third-generation cephalosporins in clinical isolates of klebsiella pneumoniae: identification of ctx-1, a novel β-lactamase," *Journal of Antimicrobial Chemotherapy*, vol. 20, no. 3, pp. 323–334, 1987.

[83] C. Kliebe, B. Nies, J. Meyer, R. Tolxdorff-Neutzling, and B. Wiedemann, "Evolution of plasmid-coded resistance to broad-spectrum cephalosporins," *Antimicrobial agents and chemotherapy*, vol. 28, no. 2, pp. 302–307, 1985.

[84] W. Sougakoff, S. Goussard, and P. Courvalin, "The tem-3 β-lactamase, which hydrolyzes broad-spectrum cephalosporins, is derived from the tem-2 penicillinase by two amino acid substitutions," *FEMS Microbiology letters*, vol. 56, no. 3, pp. 343–348, 1988.

[85] M. L. Salverda, J. A. G. De Visser, and M. Barlow, "Natural evolution of tem-1 β-lactamase: experimental reconstruction and clinical relevance," *FEMS microbiology reviews*, vol. 34, no. 6, pp. 1015–1036, 2010.

[86] D. Shcherbinin, M. Y. Rubtsova, V. Grigorenko, I. Uporov, A. Veselovsky, and A. Egorov, "The study of the role of mutations m182t and q39k in the tem-72 β-lactamase structure by the molecular dynamics method," *Biochemistry (Moscow), Supplement Series B: Biomedical Chemistry*, vol. 11, pp. 120–127, 2017.

[87] D. Carcione, C. Siracusa, A. Sulejmani, V. Leoni, and J. Intra, "Old and new beta-lactamase inhibitors: Molecular structure, mechanism of action, and clinical use," *Antibiotics*, vol. 10, no. 8, p. 995, 2021.

[88] G. G. Zhanel, C. D. Lawson, H. Adam, F. Schweizer, S. Zelenitsky, P. R. Lagacé-Wiens, A. Denisuik, E. Rubinstein, A. S. Gin, D. J. Hoban, *et al.*, "Ceftazidime-avibactam: a novel cephalosporin/β-lactamase inhibitor combination," *Drugs*, vol. 73, pp. 159–177, 2013.

[89] E. J. Zasowski, J. M. Rybak, and M. J. Rybak, "The β-lactams strike back: Ceftazidime-avibactam," *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, vol. 35, no. 8, pp. 755–770, 2015.

[90] Y. Lee, J. Kim, and S. Trinh, "Meropenem–vaborbactam (vabomereâ¢): another option for carbapenem-resistant enterobacteriaceae," *Pharmacy and Therapeutics*, vol. 44, no. 3, p. 110, 2019.

[91] T. S. Patel, J. M. Pogue, J. P. Mills, and K. S. Kaye, "Meropenem–vaborbactam: a new weapon in the war against infections due to resistant gram-negative bacteria," *Future microbiology*, vol. 13, no. 09, pp. 971–983, 2018.

[92] T. A. Campanella and J. C. Gallagher, "A clinical review and critical evaluation of imipenem-relebactam: evidence to date," *Infection and drug resistance*, pp. 4297–4308, 2020.

[93] Z. Zhang and T. Palzkill, "Dissecting the protein-protein interface between β-lactamase inhibitory protein and class a β-lactamases," *Journal of Biological Chemistry*, vol. 279, no. 41, pp. 42860–42866, 2004.

[94] M. Gretes, D. C. Lim, L. de Castro, S. E. Jensen, S. G. Kang, K. J. Lee, and N. C. Strynadka, "Insights into positive and negative requirements for protein–protein interactions by crystallographic analysis of the β-lactamase inhibitory proteins blip, blip-i, and blp," *Journal of molecular biology*, vol. 389, no. 2, pp. 289–305, 2009.

[95] N. C. Strynadka, S. E. Jensen, P. M. Alzari, and M. N. James, "A potent new mode of β-lactamase inhibition revealed by the 1.7 å x-ray crystallographic structure of the tem-1–blip complex," *Nature structural biology*, vol. 3, no. 3, pp. 290–297, 1996.

[96] H. U. Park and K. J. Lee, "Cloning and heterologous expression of the gene for blip-ii, a-lactamase-inhibitory protein from streptomyces exfoliatus smf19," *Microbiology*, vol. 144, no. 8, pp. 2161–2167, 1998.

[97] S. G. Kang, H. U. Park, H. S. Lee, H. T. Kim, and K. J. Lee, "New β-lactamase inhibitory protein (blip-i) from streptomyces exfoliatus smf19 and its roles on the morphological differentiation," *Journal of Biological Chemistry*, vol. 275, no. 22, pp. 16851–16856, 2000.

[98] W. Eiamphungporn, N. Schaduangrat, A. A. Malik, and C. Nantasenamat, "Tackling the antibiotic resistance caused by class a β-lactamases through the use of β-lactamase inhibitory protein," *International Journal of Molecular Sciences*, vol. 19, no. 8, p. 2222, 2018.

[99] K. A. Reynolds, J. M. Thomson, K. D. Corbett, C. R. Bethel, J. M. Berger, J. F. Kirsch, R. A. Bonomo, and T. M. Handel, "Structural and computational characterization of the shv-1 β-lactamase-β-lactamase inhibitor protein interface," *Journal of Biological Chemistry*, vol. 281, no. 36, pp. 26745–26753, 2006.

[100] Z. Zhang and T. Palzkill, "Determinants of binding affinity and specificity for the interaction of tem-1 and sme-1 β-lactamase with β-lactamase inhibitory

protein," *Journal of Biological Chemistry*, vol. 278, no. 46, pp. 45706–45712, 2003.

[101] F. G. Banting, C. H. Best, J. B. Collip, W. R. Campbell, and A. A. Fletcher, "Pancreatic extracts in the treatment of diabetes mellitus," *Canadian Medical Association Journal*, vol. 12, no. 3, p. 141, 1922.

[102] M. Muttenthaler, G. F. King, D. J. Adams, and P. F. Alewood, "Trends in peptide drug discovery," *Nature reviews Drug discovery*, vol. 20, no. 4, pp. 309–325, 2021.

[103] A. C.-L. Lee, J. L. Harris, K. K. Khanna, and J.-H. Hong, "A comprehensive review on current advances in peptide drug development and design," *International journal of molecular sciences*, vol. 20, no. 10, p. 2383, 2019.

[104] A. Bin Hafeez, X. Jiang, P. J. Bergen, and Y. Zhu, "Antimicrobial peptides: An update on classifications and databases," *International journal of molecular sciences*, vol. 22, no. 21, p. 11691, 2021.

[105] A. Reinhardt and I. Neundorf, "Design and application of antimicrobial peptide conjugates," *International journal of molecular sciences*, vol. 17, no. 5, p. 701, 2016.

[106] A. A. David, S. E. Park, K. Parang, and R. K. Tiwari, "Antibiotics-peptide conjugates against multidrug-resistant bacterial pathogens," *Current Topics in Medicinal Chemistry*, vol. 18, no. 22, pp. 1926–1936, 2018.

[107] S. Desgranges, C. C. Ruddle, L. P. Burke, T. M. McFadden, J. E. O'Brien, D. Fitzgerald-Hughes, H. Humphreys, T. P. Smyth, and M. Devocelle, "β-lactam-host defence peptide conjugates as antibiotic prodrug candidates targeting resistant bacteria," *Rsc Advances*, vol. 2, no. 6, pp. 2480–2492, 2012.

[108] G. W. Rudgers, W. Huang, and T. Palzkill, "Binding properties of a peptide derived from β-lactamase inhibitory protein," *Antimicrobial agents and chemotherapy*, vol. 45, no. 12, pp. 3279–3286, 2001.

[109] N. Berdigaliyev and M. Aljofan, "An overview of drug discovery and development," *Future medicinal chemistry*, vol. 12, no. 10, pp. 939–947, 2020.

[110] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Advanced drug delivery reviews*, vol. 23, no. 1-3, pp. 3–25, 1997.

[111] W. Zeng, L. Guo, S. Xu, J. Chen, and J. Zhou, "High-throughput screening technology in industrial biotechnology," *Trends in biotechnology*, vol. 38, no. 8, pp. 888–906, 2020.

[112] X. Lin, X. Li, and X. Lin, "A review on applications of computational methods in drug screening and design," *Molecules*, vol. 25, no. 6, p. 1375, 2020.

[113] A. Tropsha, "Best practices for qsar model development, validation, and exploitation," *Molecular informatics*, vol. 29, no. 6-7, pp. 476–488, 2010.

[114] K. Fosgerau and T. Hoffmann, "Peptide therapeutics: current status and future directions," *Drug discovery today*, vol. 20, no. 1, pp. 122–128, 2015.

[115] I. S. Sigal, B. G. Harwood, and R. Arentzen, "Thiol-beta-lactamase: replacement of the active-site serine of rtem beta-lactamase by a cysteine residue.," *Proceedings of the National Academy of Sciences*, vol. 79, no. 23, pp. 7157–7160, 1982.

[116] G. Winter, A. R. Fersht, A. J. Wilkinson, M. Zoller, and M. Smith, "Redesigning enzyme structure by site-directed mutagenesis: tyrosyl trna synthetase and atp binding," *Nature*, vol. 299, no. 5885, pp. 756–758, 1982.

[117] M. T. Muhammed and E. Aki-Yalcin, "Homology modeling in drug discovery: Overview, current applications, and future perspectives," *Chemical Biology amp; Drug Design*, vol. 93, no. 1, pp. 12–20, 2019.

[118] B. Kuhlman and P. Bradley, "Advances in protein structure prediction and design," *Nature Reviews Molecular Cell Biology*, vol. 20, no. 11, pp. 681–697, 2019.

[119] A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, and J. Moult, "Critical assessment of methods of protein structure prediction (casp)âround xiv," *Proteins: Structure, Function, and Bioinformatics*, vol. 89, no. 12, pp. 1607–1617, 2021.

[120] E. King, E. Aitchison, H. Li, and R. Luo, "Recent developments in free energy calculations for drug discovery," *Frontiers in Molecular Biosciences*, vol. 8, p. 712085, 2021.

[121] A. de Ruiter and C. Oostenbrink, "Advances in the calculation of binding free energies," *Current opinion in structural biology*, vol. 61, pp. 207–212, 2020.

[122] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method," *Journal of computational chemistry*, vol. 13, no. 8, pp. 1011–1021, 1992.

[123] J. Li, A. Fu, and L. Zhang, "An overview of scoring functions used for protein–ligand interactions in molecular docking," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 11, pp. 320–328, 2019.

[124] L. Hu, X. Wang, Y.-A. Huang, P. Hu, and Z.-H. You, "A survey on computational models for predicting protein–protein interactions," *Briefings in bioinformatics*, vol. 22, no. 5, p. bbab036, 2021.

[125] T. Siebenmorgen and M. Zacharias, "Computational prediction of protein–protein binding affinities," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 10, no. 3, p. e1448, 2020.

[126] D. Kozakov, D. R. Hall, B. Xia, K. A. Porter, D. Padhorny, C. Yueh, D. Beglov, and S. Vajda, "The cluspro web server for protein–protein docking," *Nature protocols*, vol. 12, no. 2, pp. 255–278, 2017.

[127] C. Dominguez, R. Boelens, and A. M. Bonvin, "Haddock: a protein- protein docking approach based on biochemical or biophysical information," *Journal of the American Chemical Society*, vol. 125, no. 7, pp. 1731–1737, 2003.

[128] M. Ohue, T. Shimoda, S. Suzuki, Y. Matsuzaki, T. Ishida, and Y. Akiyama, "Megadock 4.0: an ultra–high-performance protein–protein docking software for heterogeneous supercomputers," *Bioinformatics*, vol. 30, no. 22, pp. 3281–3283, 2014.

[129] R. Chen, L. Li, and Z. Weng, "Zdock: an initial-stage protein-docking algorithm," *Proteins: Structure, Function, and Bioinformatics*, vol. 52, no. 1, pp. 80–87, 2003.

[130] S. A. Hollingsworth and R. O. Dror, "Molecular dynamics simulation for all," *Neuron*, vol. 99, no. 6, pp. 1129–1143, 2018.

[131] P. Dauber-Osguthorpe and A. T. Hagler, "Biomolecular force fields: where have we been, where are we now, where do we need to go and how do we get there?," *Journal of computer-aided molecular design*, vol. 33, no. 2, pp. 133–203, 2019.

[132] I. Poltavsky and A. Tkatchenko, "Machine learning force fields: Recent advances and remaining challenges," *The Journal of Physical Chemistry Letters*, vol. 12, no. 28, pp. 6551–6564, 2021.

[133] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen, "Gromacs: fast, flexible, and free," *Journal of computational chemistry*, vol. 26, no. 16, pp. 1701–1718, 2005.

[134] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, vol. 1, pp. 19–25, 2015.

[135] W. Huang, Q.-Q. Le, M. LaRocco, and T. Palzkill, "Effect of threonine-to-methionine substitution at position 265 on structure and function of tem-1 beta-lactamase," *Antimicrobial agents and chemotherapy*, vol. 38, no. 10, pp. 2266–2269, 1994.

[136] S. Ohsuka, Y. Arakawa, T. Horii, H. Ito, and M. Ohta, "Effect of ph on activities of novel beta-lactamases and beta-lactamase inhibitors against these beta-lactamases," *Antimicrobial agents and chemotherapy*, vol. 39, no. 8, pp. 1856–1858, 1995.

[137] V. Stojanoski, D.-C. Chow, L. Hu, B. Sankaran, H. F. Gilbert, B. V. Prasad, and T. Palzkill, "A triple mutant in the ω-loop of tem-1 β-lactamase changes the substrate profile via a large conformational change and an altered general base for catalysis," *Journal of Biological Chemistry*, vol. 290, no. 16, pp. 10382–10394, 2015.

[138] J. P. Quinn, D. Miyashiro, D. Sahm, R. Flamm, and K. Bush, "Novel plasmid-mediated beta-lactamase (tem-10) conferring selective resistance to ceftazidime and aztreonam in clinical isolates of klebsiella pneumoniae," *Antimicrobial agents and chemotherapy*, vol. 33, no. 9, pp. 1451–1456, 1989.

[139] C. Poyart, P. Mugnier, G. Quesne, P. Berche, and P. Trieu-Cuot, "A novel extended-spectrum tem-type β-lactamase (tem-52) associated with decreased susceptibility to moxalactam in klebsiella pneumoniae," *Antimicrobial agents and chemotherapy*, vol. 42, no. 1, pp. 108–113, 1998.

[140] J. G. Sutcliffe, "Nucleotide sequence of the ampicillin resistance gene of escherichia coli plasmid pbr322.," *Proceedings of the National Academy of Sciences*, vol. 75, no. 8, pp. 3737–3741, 1978.

[141] W. L. DeLano *et al.*, "Pymol: An open-source molecular graphics tool," *CCP4 Newsl. Protein Crystallogr*, vol. 40, no. 1, pp. 82–92, 2002.

[142] S. Ness, R. Martin, A. M. Kindler, M. Paetzel, M. Gold, S. E. Jensen, J. B. Jones, and N. C. Strynadka, "Structure-based design guides the improved efficacy of deacylation transition state analogue inhibitors of tem-1 β-lactamase," *Biochemistry*, vol. 39, no. 18, pp. 5312–5321, 2000.

[143] M. C. Orencia, J. S. Yoon, J. E. Ness, W. P. Stemmer, and R. C. Stevens, "Predicting the emergence of antibiotic resistance by directed evolution and structural analysis," *Nature structural biology*, vol. 8, no. 3, pp. 238–242, 2001.

[144] X. Wang, G. Minasov, and B. K. Shoichet, "Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs," *Journal of molecular biology*, vol. 320, no. 1, pp. 85–95, 2002.

[145] X. Wang, G. Minasov, and B. K. Shoichet, "The structural bases of antibiotic resistance in the clinically derived mutant β-lactamases tem-30, tem-32, and tem-34," *Journal of Biological Chemistry*, vol. 277, no. 35, pp. 32149–32156, 2002.

[146] E. Fonzé, P. Charlier, Y. To'Th, M. Vermeire, X. Raquet, A. Dubus, and J.-M. Frere, "Tem1 β-lactamase structure solved by molecular replacement and refined

structure of the s235a mutant," *Acta Crystallographica Section D: Biological Crystallography*, vol. 51, no. 5, pp. 682–694, 1995.

[147] N. G. Brown, S. Shanker, B. V. Prasad, and T. Palzkill, "Structural and biochemical evidence that a tem-1 β-lactamase n170g active site mutant acts via substrate-assisted catalysis," *Journal of Biological Chemistry*, vol. 284, no. 48, pp. 33703–33712, 2009.

[148] M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek, and G. R. Hutchison, "Avogadro: an advanced semantic chemical editor, visualization, and analysis platform," *Journal of cheminformatics*, vol. 4, no. 1, pp. 1–17, 2012.

[149] F. Wang, L. Shen, H. Zhou, S. Wang, X. Wang, and P. Tao, "Machine learning classification model for functional binding modes of tem-1 β-lactamase," *Frontiers in molecular biosciences*, vol. 6, p. 47, 2019.

[150] Y.-M. Chang, W.-Y. Jeng, T.-P. Ko, Y.-J. Yeh, C. K.-M. Chen, and A. H.-J. Wang, "Structural study of tcar and its complexes with multiple antibiotics from staphylococcus epidermidis," *Proceedings of the National Academy of Sciences*, vol. 107, no. 19, pp. 8617–8622, 2010.

[151] D. Bellini, L. Koekemoer, H. Newman, and C. G. Dowson, "Novel and improved crystal structures of h. influenzae, e. coli and p. aeruginosa penicillin-binding protein 3 (pbp3) and n. gonorrhoeae pbp2: toward a better understanding of β-lactam target-mediated resistance," *Journal of molecular biology*, vol. 431, no. 18, pp. 3501–3519, 2019.

[152] C. J. Adamski, A. M. Cardenas, N. G. Brown, L. B. Horton, B. Sankaran, B. V. Prasad, H. F. Gilbert, and T. Palzkill, "Molecular basis for the catalytic specificity of the ctx-m extended-spectrum β-lactamases," *Biochemistry*, vol. 54, no. 2, pp. 447–457, 2015.

[153] M. P. Patel, L. Hu, V. Stojanoski, B. Sankaran, B. V. Prasad, and T. Palzkill, "The drug-resistant variant p167s expands the substrate profile of ctx-m β-lactamases for oxyimino-cephalosporin antibiotics by enlarging the active site upon acylation," *Biochemistry*, vol. 56, no. 27, pp. 3443–3453, 2017.

[154] O. Trott and A. J. Olson, "Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *Journal of computational chemistry*, vol. 31, no. 2, pp. 455–461, 2010.

[155] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the amber ff99sb protein force field," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 8, pp. 1950–1958, 2010.

[156] J. Wang, W. Wang, P. A. Kollman, and D. A. Case, "Antechamber: an accessory software package for molecular mechanical calculations," *J. Am. Chem. Soc*, vol. 222, no. 1, 2001.

[157] A. W. Sousa da Silva and W. F. Vranken, "Acpype-antechamber python parser interface," *BMC research notes*, vol. 5, pp. 1–8, 2012.

[158] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," *The Journal of chemical physics*, vol. 126, no. 1, 2007.

[159] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," *Journal of Applied physics*, vol. 52, no. 12, pp. 7182–7190, 1981.

[160] H. Gray, G. G. Leday, C. A. Vallejos, and S. Richardson, "Shrinkage estimation of large covariance matrices using multiple shrinkage targets," *arXiv preprint arXiv:1809.08024*, 2018.

[161] H. Gunawan, O. Neswan, and W. Setya-Budhi, "A formula for angles between subspaces of inner product spaces.," *Beiträge zur Algebra und Geometrie*, vol. 46, no. 2, pp. 311–320, 2005.

[162] J. Miao and A. Ben-Israel, "On principal angles between subspaces in rn," *Linear algebra and its applications*, vol. 171, pp. 81–98, 1992.

[163] A. Amadei, M. A. Ceruso, and A. Di Nola, "On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations," *Proteins: Structure, Function, and Bioinformatics*, vol. 36, no. 4, pp. 419–424, 1999.

[164] A. Bunse-Gerstner, R. Byers, and V. Mehrmann, "Numerical methods for simultaneous diagonalization," *SIAM journal on matrix analysis and applications*, vol. 14, no. 4, pp. 927–949, 1993.

[165] J. Patterson, C. Avery, T. Grear, and D. J. Jacobs, "Biased hypothesis formation from projection pursuit," *arXiv preprint arXiv:2201.00889*, 2022.

[166] T. Grear and D. Jacobs, "Classifying eeg motor imagery signals using supervised projection pursuit for artefact removal," in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2952–2958, IEEE, 2021.

[167] J. R. Horn and B. K. Shoichet, "Allosteric inhibition through core disruption," *Journal of molecular biology*, vol. 336, no. 5, pp. 1283–1291, 2004.

[168] N. Doucet, P.-Y. De Wals, and J. N. Pelletier, "Site-saturation mutagenesis of tyr-105 reveals its importance in substrate stabilization and discrimination in tem-1 β-lactamase," *Journal of Biological Chemistry*, vol. 279, no. 44, pp. 46295–46303, 2004.

[169] N. Doucet, P.-Y. Savard, J. N. Pelletier, *et al.*, "Nmr investigation of tyr105 mutants in tem-1 β-lactamase: dynamics are correlated with function," *Journal of Biological Chemistry*, vol. 282, no. 29, pp. 21448–21459, 2007.

[170] E. Hellemann, A. Nallathambi, and J. D. Durrant, "Allosteric inhibition of tem-1 β lactamase: Microsecond molecular dynamics simulations provide mechanistic insights," *Protein Science*, vol. 32, no. 4, p. e4622, 2023.

[171] T. Palzkill, "Structural and mechanistic basis for extended-spectrum drug-resistance mutations in altering the specificity of tem, ctx-m, and kpc β-lactamases," *Frontiers in molecular biosciences*, vol. 5, p. 16, 2018.

[172] A. Huletsky, J. Knox, and R. Levesque, "Role of ser-238 and lys-240 in the hydrolysis of third-generation cephalosporins by shv-type beta-lactamases probed by site-directed mutagenesis and three-dimensional modeling.," *Journal of biological chemistry*, vol. 268, no. 5, pp. 3690–3697, 1993.

[173] C. Cantu and T. Palzkill, "The role of residue 238 of tem-1 β-lactamase in the hydrolysis of extended-spectrum antibiotics," *Journal of Biological Chemistry*, vol. 273, no. 41, pp. 26603–26609, 1998.

[174] S. B. Vakulenko, P. Taibi-Tronche, M. Tóth, I. Massova, S. A. Lerner, and S. Mobashery, "Effects on substrate profile by mutational substitutions at positions 164 and 179 of the class a tempuc19 β-lactamase from escherichia coli," *Journal of Biological Chemistry*, vol. 274, no. 33, pp. 23052–23060, 1999.

[175] V. Agarwal, T. C. Yadav, A. Tiwari, and P. Varadwaj, "Detailed investigation of catalytically important residues of class a β-lactamase," *Journal of Biomolecular Structure and Dynamics*, vol. 41, no. 5, pp. 2046–2073, 2023.

[176] O. A. Pemberton, R. E. Noor, V. Kumar MV, R. Sanishvili, M. T. Kemp, F. L. Kearns, H. L. Woodcock, I. Gelis, and Y. Chen, "Mechanism of proton transfer in class a β-lactamase catalysis and inhibition by avibactam," *Proceedings of the National Academy of Sciences*, vol. 117, no. 11, pp. 5818–5825, 2020.

[177] I. Kather, R. P. Jakob, H. Dobbek, and F. X. Schmid, "Increased folding stability of tem-1 β-lactamase by in vitro selection," *Journal of molecular biology*, vol. 383, no. 1, pp. 238–251, 2008.

[178] D. C. Marciano, J. M. Pennington, X. Wang, J. Wang, Y. Chen, V. L. Thomas, B. K. Shoichet, and T. Palzkill, "Genetic and structural characterization of an l201p global suppressor substitution in tem-1 β-lactamase," *Journal of molecular biology*, vol. 384, no. 1, pp. 151–164, 2008.

[179] I. Galdadas, S. Qu, A. S. F. Oliveira, E. Olehnovics, A. R. Mack, M. F. Mojica, P. K. Agarwal, C. L. Tooke, F. L. Gervasio, J. Spencer, *et al.*, "Allosteric communication in class a β-lactamases occurs via cooperative coupling of loop dynamics," *Elife*, vol. 10, p. e66567, 2021.

[180] S. G. Tristram, "Effect of extended-spectrum β-lactamases on the susceptibility of haemophilus influenzae to cephalosporins," *Journal of Antimicrobial Chemotherapy*, vol. 51, no. 1, pp. 39–43, 2003.

[181] F.-X. Weill, M. Demartin, L. Fabre, and P. A. Grimont, "Extended-spectrum-β-lactamase (tem-52)-producing strains of salmonella enterica of various serotypes isolated in france," *Journal of clinical microbiology*, vol. 42, no. 7, pp. 3359–3362, 2004.

[182] G. A. Jacoby and A. A. Medeiros, "More extended-spectrum beta-lactamases," *Antimicrobial agents and chemotherapy*, vol. 35, no. 9, pp. 1697–1704, 1991.

[183] O. Fisette, S. Morin, P.-Y. Savard, P. Lagüe, and S. M. Gagné, "Tem-1 backbone dynamicsâinsights from combined molecular dynamics and nuclear magnetic resonance," *Biophysical journal*, vol. 98, no. 4, pp. 637–645, 2010.

[184] S. J. Marrink, L. Monticelli, M. N. Melo, R. Alessandri, D. P. Tieleman, and P. C. Souza, "Two decades of martini: Better beads, broader scope," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 13, no. 1, p. e1620, 2023.

[185] X. Periole and S.-J. Marrink, "The martini coarse-grained force field," *Biomolecular simulations: Methods and protocols*, pp. 533–565, 2013.

[186] J. Farmer, F. Kanwal, N. Nikulsin, M. C. Tsilimigras, and D. J. Jacobs, "Statistical measures to quantify similarity between molecular dynamics simulation trajectories," *Entropy*, vol. 19, no. 12, p. 646, 2017.

[187] S. Genheden and U. Ryde, "Will molecular dynamics simulations of proteins ever reach equilibrium?," *Physical Chemistry Chemical Physics*, vol. 14, no. 24, pp. 8662–8677, 2012.

[188] K. M. Papp-Wallace, A. R. Mack, M. A. Taracila, and R. A. Bonomo, "Resistance to novel β-lactam–β-lactamase inhibitor combinations: the âprice of progressâ," *Infectious Disease Clinics*, vol. 34, no. 4, pp. 773–819, 2020.

[189] X. Wang, D. Ni, Y. Liu, and S. Lu, "Rational design of peptide-based inhibitors disrupting protein-protein interactions," *Frontiers in chemistry*, vol. 9, p. 682675, 2021.

[190] W. M. Dawson, G. G. Rhys, and D. N. Woolfson, "Towards functional de novo designed proteins," *Current Opinion in Chemical Biology*, vol. 52, pp. 102–111, 2019.

[191] D. N. Woolfson, "A brief history of de novo protein design: minimal, rational, and computational," *Journal of Molecular Biology*, vol. 433, no. 20, p. 167160, 2021.

[192] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic acids research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[193] M. Bhagwat and L. Aravind, "Psi-blast tutorial," *Comparative genomics*, pp. 177–186, 2008.

[194] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang, "The i-tasser suite: protein structure and function prediction," *Nature methods*, vol. 12, no. 1, pp. 7–8, 2015.

[195] R. Wu, F. Ding, R. Wang, R. Shen, X. Zhang, S. Luo, C. Su, Z. Wu, Q. Xie, B. Berger, *et al.*, "High-resolution de novo structure prediction from primary sequence," *BioRxiv*, pp. 2022–07, 2022.

[196] S. Wells, S. Menor, B. Hespenheide, and M. F. Thorpe, "Constrained geometric simulation of diffusive motion in proteins," *Physical Biology*, vol. 2, no. 4, p. S127, 2005.

[197] D. J. Jacobs, L. A. Kuhn, and M. F. Thorpe, "Flexible and rigid regions in proteins," *Rigidity theory and applications*, pp. 357–384, 2002.

[198] A. D. Vogt and E. Di Cera, "Conformational selection is a dominant mechanism of ligand binding," *Biochemistry*, vol. 52, no. 34, pp. 5723–5729, 2013.

[199] M. D. Balcewich, T. M. Reeve, E. A. Orlikow, L. J. Donald, D. J. Vocadlo, and B. L. Mark, "Crystal structure of the ampr effector binding domain provides insight into the molecular regulation of inducible ampc β-lactamase," *Journal of molecular biology*, vol. 400, no. 5, pp. 998–1010, 2010.

[200] J.-H. Shin, A. G. Sulpizio, A. Kelley, L. Alvarez, S. G. Murphy, L. Fan, F. Cava, Y. Mao, M. A. Saper, and T. Dörr, "Structural basis of peptidoglycan endopeptidase regulation," *Proceedings of the National Academy of Sciences*, vol. 117, no. 21, pp. 11692–11702, 2020.

[201] R. Cohen-Khait, O. Dym, S. Hamer-Rogotner, and G. Schreiber, "Promiscuous protein binding as a function of protein stability," *Structure*, vol. 25, no. 12, pp. 1867–1874, 2017.

[202] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, *et al.*, "Biopython: freely available python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, p. 1422, 2009.

[203] T. Bepler and B. Berger, "Learning the protein language: Evolution, structure, and function," *Cell systems*, vol. 12, no. 6, pp. 654–669, 2021.

[204] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, *et al.*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, p. e2016239118, 2021.

[205] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[206] S. Deorowicz, A. Debudaj-Grabysz, and A. Gudyś, "Famsa: Fast and accurate multiple sequence alignment of huge protein families," *Scientific reports*, vol. 6, no. 1, p. 33964, 2016.

[207] S. R. Eddy, "Where did the blosum62 alignment score matrix come from?," *Nature biotechnology*, vol. 22, no. 8, pp. 1035–1036, 2004.

[208] M. Startek, S. Lasota, M. Sykulski, A. Bułak, L. Noé, G. Kucherov, and A. Gambin, "Efficient alternatives to psi-blast," *Bulletin of the Polish Academy of Sciences. Technical Sciences*, vol. 60, no. 3, pp. 495–505, 2012.

[209] R. Evans, M. OâNeill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, *et al.*, "Protein complex prediction with alphafold-multimer," *biorxiv*, pp. 2021–10, 2021.

[210] E. F. McDonald, T. Jones, L. Plate, J. Meiler, and A. Gulsevin, "Benchmarking alphafold2 on peptide structure prediction," *Structure*, vol. 31, no. 1, pp. 111–119, 2023.

[211] C. J. Wilson, W.-Y. Choy, and M. Karttunen, "Alphafold2: a role for disordered protein/region prediction?," *International Journal of Molecular Sciences*, vol. 23, no. 9, p. 4591, 2022.

[212] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, "Colabfold: making protein folding accessible to all," *Nature methods*, vol. 19, no. 6, pp. 679–682, 2022.

[213] A. V. Finkelstein, "Does alphafold predict the spatial structure of a protein from physics or recognize it (its main parts and their association) using databases?," *bioRxiv*, pp. 2022–11, 2022.

[214] C. Outeiral, D. A. Nissley, and C. M. Deane, "Current structure predictors are not learning the physics of protein folding," *Bioinformatics*, vol. 38, no. 7, pp. 1881–1887, 2022.

[215] J. M. Michaud, A. Madani, and J. S. Fraser, "A language model beats alphafold2 on orphans," *Nature Biotechnology*, vol. 40, no. 11, pp. 1576–1577, 2022.

[216] O. Lichtarge, H. R. Bourne, and F. E. Cohen, "An evolutionary trace method defines binding surfaces common to protein families," *Journal of molecular biology*, vol. 257, no. 2, pp. 342–358, 1996.