

Clemson University

TigerPrints

All Dissertations

Dissertations

8-2023

The Influence of Allostery Governing the Changes in Protein Dynamics Upon Substitution

Joseph Hess
hess7@clemson.edu

Follow this and additional works at: https://tigerprints.clemson.edu/all_dissertations

 Part of the [Biological and Chemical Physics Commons](#), [Computational Chemistry Commons](#), [Data Science Commons](#), and the [Physical Chemistry Commons](#)

Recommended Citation

Hess, Joseph, "The Influence of Allostery Governing the Changes in Protein Dynamics Upon Substitution" (2023). *All Dissertations*. 3254.
https://tigerprints.clemson.edu/all_dissertations/3254

This Dissertation is brought to you for free and open access by the Dissertations at TigerPrints. It has been accepted for inclusion in All Dissertations by an authorized administrator of TigerPrints. For more information, please contact kokeefe@clemson.edu.

THE INFLUENCE OF ALLOSTERY GOVERNING THE CHANGES IN PROTEIN
DYNAMICS UPON SUBSTITUTION

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Chemistry

by
Joseph Hess
August 2023

Accepted by:
Dr Brian Dominy, Committee Chair
Dr Steve Stuart
Dr Andrew Brown
Dr William Richardson

ABSTRACT

The focus of this research is to investigate the effects of allostery on the function/activity of an enzyme, human immunodeficiency virus type 1 (HIV-1) protease, using well-defined statistical analyses of the dynamic changes of the protein and variants with unique single point substitutions¹. The experimental data¹ evaluated here only characterized HIV-1 protease with one of its potential target substrates. Probing the dynamic interactions of the residues of an enzyme and its variants can offer insight of the developmental importance for allosteric signaling and their connection to a protein's function. The realignment of the secondary structure elements can modulate the mobility along with the frequency of residue contacts as well as which residues are making contact together²⁻⁵. We postulate that if there are more contacts occurring within a structure the mobility is being constrained and therefore gaining novel contacts can negatively influence the function of a protein.

The evolutionary importance of protein dynamics is probed by analyzing the residue positions possessing significant correlations and the relationship between experimental information¹ (variant activities). We propose that the correlated dynamics of residues observed to have considerable correlations, if disrupted, can be used to infer the function of HIV-1 protease and its variants. Given the robustness of HIV-1 protease the identification of any significant constraint imposed on the dynamics from a potential allosteric site found to disrupt the catalytic activity of the variant is not plainly evident. We

also develop machine learning (ML) algorithms to predict the protein function/activity change caused by a single point substitution by using the DCC of each residue pair. Recognition of any substantial association between the dynamics of specific residues and allosteric communication or mechanism requires detailed examination of the dynamics of HIV-1 protease and its variants.

We also explore the non-linear dependency between each pair of residues using Mutual Information (MI) and how it can influence the dynamics of HIV-1 protease and its variants. We suggest that if the residues of a protein receive more or less information than that of the WT it will adversely impact the function of the protein and can be used to support the classification of a variant structure. Furthermore, using the MI of residues obtained from the MD simulations for the HIV-1 protease structure, we build a ML model to predict a protein's change in function caused by a single point substitution. Effectively the mobility, dynamics, and non-linear features tested in these studies are found to be useful towards the prediction of potentially drug resistant substitutions related to the catalytic efficiency of HIV-1 protease and the variants.

DEDICATION

My work is dedicated to my loving wife and sons who have been a constant source of support and motivation. A special thank you to my parents for all they have given me, as well as a to my wife for her dedication and compassion.

ACKNOWLEDGMENTS

I would like to thank Dr Brian Dominy for his guidance and valuable insight he gave me to challenge myself. I am also grateful for the support and guidance my committee members Dr Steve Stuart, Dr Andrew Brown, and Dr William Richardson offered me towards my research. I would also like to thank the Chemistry Department at Clemson University for awarding me this opportunity. I would like to thank CCIT and the Palmetto crew for their work and allotting the use of the Palmetto cluster.

TABLE OF CONTENTS

	Page
TITLE PAGE	i
ABSTRACT	ii
DEDICATION	iv
ACKNOWLEDGMENTS.....	v
LIST OF TABLES	ix
LIST OF FIGURES.....	x
CHAPTER	
I. INTRODUCTION	1
Protein Fundamentals	1
Molecular Dynamics	7
Contacts	8
Stability	10
Dynamic Cross Correlation	13
Mutual Information	14
HIV-1 Protease	16

II. CHANGES IN THE MOBILITY AND DYNAMICS WHICH IMPOSE CONSTRAINTS ON THE FUNCTION OF A PROTEIN.....	18
2.1 Introduction	18
2.2 Methods	27
2.3 Energy	30
2.4 Chemical & Structural Properties of HIV-1 Protease	32
2.5 Results and Discussion.....	35
2.6 Conclusion.....	53
III. THE IMPORTANCE OF ANTICORRELATION ON THE FUNCTION OF HIV-1 PROTEASE VARIANTS AND PREDICTING THE FUNCTION OF THE PROTEIN UPON SINGLE POINT SUBSTITUTION USING A ROBUST MACHINE.....	55
3.1 Introduction	55
3.2 Methods	61
3.2 Machine Learning	65
3.3 Results and Discussion.....	67
3.4 Conclusion.....	81
IV. UNDERSTANDING THE ROLE OF DYNAMICS USING MUTUAL INFORMATION AND DEVELOPING A ROBUST MACHINE LEARNING MODEL TO PREDICT THE FUNCTION OF HIV-1 PROTEASE VARIANTS.....	84
4.1 Introduction	84
4.2 Methods	90
4.3 Machine Learning Model	95

4.3 Results and Discussion.....	97
4.4 Conclusion.....	107
APPENDICES.....	110
A: Supplementary Information Chapter II	111
B: Supplementary Information Chapter III.....	118
C: Supplementary Information Chapter IV	128
D: Challenges and Future work.....	134
REFERENCES.....	136

LIST OF TABLES

Table	Page
2.2 Table 2.1 Comparison of the Chemical and Physical Features for the WT and each of the 12 CA and 56 CI variants (y-axis) with the 99 amino acids of the HIV-1 protease (x-axis)	32
2.2 Table 2.2 Change in folding free energy (dG) and change in folding free energy upon substitution (ddG) using MMPBSA for method 1 (average) and method 2 (difference).....	49
2.2 Table 2.3 Change in folding free energy upon substitution (ddG) using the DUET, INPS, Eris, and DDGun webserver ^{105, 110, 227-235}	51
2.2 Table 3.1 Group 1 corresponds to the positions showing strongly negative correlation and Group 2 corresponds to the positions showing relatively weak negative correlation	77
2.1 Table 3.2 8 runs of the XGBoost model using different random seeds	81
2.2 Table 4.1 8 runs of the logistic regression model using different random seeds	107

LIST OF FIGURES

Figure	Page
1.1	Figure 1.1 Shows the A) basic amino acid structure, B) amino acid chain formation, C) dihedral angle, and D) secondary structure elements2
1.2	Figure 1.2 Reaction diagram for the folding free energy change (A) and binding free energy change (B) of a protein structure. Figure 1.2 (A) shows the three steps of folding of a protein; (1) unfolded monomer, (2) the formation of secondary structure elements, and (3) the completely folded monomer. Figure 1.2 (B) shows the three steps associated with the binding and catalysis of a substrate; enzyme (E) + substrate (S), the ES complex, and the E + product (P)..... 12
2.1	Figure 2.1 The figure shown above is a front and back view of the HIV-1 protease (PDB id: 1hxx). The colored mutant residue positions are highlighted in a single chain, though each substitution was applied to both chains, to show the location of the CI and CA variant residue positions (red and green respectively). Of the 99 common residues that make up each chain, there are 42 unique positions are found to cause an inactive variant and 12 positions are found to cause activity equal or greater than the WT structure ¹ . There are two residue positions that have both zero and greater activity, residues K20 and I64. The K20 CA variant residue is K20R, a basic-to-basic side chain substitution, but the K20 CI variant residue is K20E, a basic to acidic side chain substitution. Similarly, for I64 CA variant residue is I64V, a non-polar to non-polar side chain substitution, whereas the CI variant residue is I64T, a non-polar to polar side chain substitution. There are also a number of redundant positions that are shown to have zero activity ¹22
2.2	Figure 2.2 The figures above show the average RMSF of the residue for chain A (y axis) and chain B (x axis) for (Figure 2.2 A) and chain B (Figure 2.2 B) of the total 72 ns of MD production for each trajectory.

List of Figures (Continued)

	The mean of all 3 trajectories is represented as the points with standard deviation (gray) and standard error (black) the best fit line shown by the equation in the top left.	36
2.2	Figure 2.3 A&B: The figures above show the RMSF (y axis) for the CA variant (left) and CI variant (right) of the residue numbers (x axis) for the total length of production. The WT is represented as the black line and each variant an assigned is shown as gray (top left)	38
2.2	Figure 2.4 A&B: The figures above show the total number of contacts that are observed in chain A for the CA variants (A) and CI variants (B). Each of the 99 residue numbers are shown on the x and y axis. The points represent contacts between 2 residues with the size of the point corresponding to the frequency observed throughout the production for the specific variant (see color legend)	39
2.2	Figure 2.5 A&B: The figures above show the most common residue contacts that are gained or lost in chain A for at least 70% of the CA variants (A) and CI variants (B). Each of the 99 residue numbers are shown on the x and y axis. The points represent contacts between 2 residues with the size of the black points corresponding to the frequency observed throughout the production for the WT. Residues that are observed in the WT but not in the variant are considered a loss (red), inversely residues observed in the variant and not the WT are considered gain (blue). The size of the points is also scaled according to the average frequency observed from the corresponding variant case.....	41
2.2	Figure 2.6 These figures identify the structural orientation of the residues found to have novel gain of contact for the CI variants	44
2.2	Figure 2.7 These figures identify the structural orientation of the residues found to have novel loss of contact for	

List of Figures (Continued)

	the CA variants	45
2.2	Figure 2.8 These figures identify the structural orientation of the residues found to have novel loss of contact for the CI variants	46
2.1	Figure 3.1 The figure shown above is a front and back view of the HIV-1 protease (PDB id: 1hwx). The colored mutant residue positions are highlighted in a single chain, though each substitution was applied to both chains, to show the location of the CI and CA variant residue positions (red and green respectively). Of the 99 common residues that make up each chain, there are 42 unique positions are found to cause an inactive variant and 12 positions are found to cause activity equal or greater than the WT structure ¹ . There are two residue positions that have both zero and greater activity, residues K20 and I64. The K20 CA variant residue is K20R, a basic-to-basic side chain substitution, but the K20 CI variant residue is K20E, a basic to acidic side chain substitution. Similarly, for I64 CA variant residue is I64V, a non-polar to non-polar side chain substitution, whereas the CI variant residue is I64T, a non-polar to polar side chain substitution. There are also a number of redundant positions that are shown to have zero activity ¹	59
2.2	Figure 3.2 (A) Representation Covariance matrix for the motion of C α atoms in WT HIV-1 protease bound to MA/CA substrate. The x and y axis are the residue numbers and the points represent a Pearson correlation value between 1 and -1 based on the average cross correlation value for 3 MD simulations of the WT structure. (B) Shows the regions of intra-protein and domain-domain correlations found in Figure 2(A)	68
2.2	Figure 3.3 Evaluation of the degradation of the minimum anticorrelation observed of the WT structure for every 10ns of production. The total production of the WT consists of 72ns and the covariance was calculated	

List of Figures (Continued)

- for each of the 6 productions above by removing 10ns from the beginning of each subsequent productions of the WT. The residue pairs that are observed to have the greatest anticorrelation are shown below the respective production 69
- 2.2 Figure 3.4 This figure examines the relationship between the degree of negatively correlated motion associated with each unique residue position in HIV-1 protease and the impact of a substitution resulting in an inactive variant toward the MA/CA substrate. If anticorrelated motion is strongly linked to enzyme activity in this system, one might expect to observe a positive slope in this figure suggesting that specific sequence positions exhibiting stronger correlated motions would also be more likely to eliminate catalytic activity upon substitution. The data here demonstrates, though there is not a positive slope, that the substitutions that result in an inactive enzyme strongly influence anticorrelation. The number of residues in each chain were collected within an iteration of every 0.03 units from -0.18 to the maximum negative correlation (-0.43). Then, the sequence positions that when mutated result in the protein having no noticeable catalytic activity (44 unique residue positions in each chain of the homodimer) are divided by all residue positions (99 residue positions in each chain of the homodimer) making up the fraction of catalytically inactive (CI) residue positions. The standard error was calculated from the confidence interval of the proportion, $z_{0.95} \sqrt{(p(1 - p))/n_p}$, shown by the T's 71
- 2.2 Figure 3.5 This figure examines the relationship between the degree of positively correlated motion associated with specific sequence positions in HIV-1 protease and the impact of a substitution of those residue positions on the peptidase activity toward the MA/CA substrate. If correlated motion were strongly linked to enzyme activity in this system, one might expect to observe a positive slope in this figure suggesting that specific sequence positions exhibiting stronger correlated motions would also be more likely to eliminate catalytic activity upon substitution. The

List of Figures (Continued)

data here demonstrates significant sensitivity of the degree of correlated motion with respect to the impact of substitution on enzyme catalytic activity. The number of unique residues were collected by iterating of every 0.07 units from 0.10 to maximum correlated motion not equal to 1. Within each bin the residue positions known to result in the protein having no noticeable catalytic activity (44 residue positions in each chain of the homodimer) are divided by all residue positions identified as having correlation in that range, making up the fraction of residue positions. The standard error was calculated from the confidence interval of the proportion, $z_{0.95} \sqrt{(p(1 - p))/n_p}$, shown by the T's 73

- 2.2 Figure 3.6A This figure examines the relationship between the degree of negatively correlated motion associated with specific sequence positions in HIV-1 protease and the impact of the average peptidase activity toward the MA/CA substrate. If correlated motion were strongly linked to enzyme activity in this system, it can be expected to observe a negative slope in this figure suggesting that specific sequence positions exhibiting stronger anticorrelated motions would also be more likely to eliminate catalytic activity upon substitution. The data here demonstrates that the degree of anticorrelated motion with respect to the impact of substitution on enzyme catalytic activity is statistically significant. The number of residues in each chain were collected within an iteration of every 0.03 units from -0.1 to the maximum negative correlation (-0.43). Then, the average of the known catalytic activity was calculated for the sequence positions within each bin. The standard error was calculated using $SE = \sqrt{\frac{(x - \mu)^2}{N}} / \sqrt{N}$, shown by the T's 75

- 2.2 Figure 3.6B This figure examines the relationship between the degree of positively correlated motion associated with specific sequence positions in HIV-1 protease and the impact of the average peptidase

List of Figures (Continued)

- activity toward the MA/CA substrate. If correlated motions were strongly linked to enzyme activity in this system, it would be expected to observe a negative slope in this figure suggesting that specific sequence positions exhibiting stronger correlated motions would also be more likely to eliminate catalytic activity upon substitution. The number of residues in each chain were collected within an iteration of every 0.07 units from 0.10 to the maximum correlation less than 1. The average of the known catalytic activity was calculated for the sequence positions within each bin. The standard error was calculated using $SE = \sqrt{\frac{(x - \mu)^2}{N}} / \sqrt{N}$, shown by the T's 76
- 2.2 Figure 3.7 Confusion matrices for the XGBoost model performance A) is the performance result of the training set and B) is the performance result of the test set with the recall, precision, and F1-score shown at the bottom right 80
- 2.1 Figure 4.1 The figure shown above is a front and back view of the HIV-1 protease (PDB id: 1hwx). The colored mutant residue positions are highlighted in a single chain, though each substitution was applied to both chains, to show the location of the CI and CA variant residue positions (red and green respectively). Of the 99 common residues that make up each chain, there are 42 unique positions are found to cause an inactive variant and 12 positions are found to cause activity equal or greater than the WT structure¹. There are two residue positions that have both zero and greater activity, residues K20 and I64. The K20 CA variant residue is K20R, a basic-to-basic side chain substitution, but the K20 CI variant residue is K20E, a basic to acidic side chain substitution. Similarly, for I64 CA variant residue is I64V, a non-polar to non-polar side chain substitution, whereas the CI variant residue is I64T, a non-polar to polar side chain substitution. There are also a number of redundant positions that are shown to have zero activity¹ 88
- 2.2 Figure 4.2 Vein diagram showing the separation of low (blue) , medium (green) and high (orange/red) MI..... 95

2.2	Figure 4.3 (A) Representation of the pairwise MI matrix of C α atoms in WT of HIV-1 protease bound to MA/CA substrate. The x and y axis are the residue numbers and the points represent the normalized MI value between 0 and 1. (B) Shows the regions of intra-domain and inter-domain correlations found in Figure 1(A).....	97
2.2	Figure 4.4 Representation of a single chain of HIV-1 protease highlighting the position of residues 34 and 35 with respect to the alpha-helix	98
2.2	Figure 4.5 Total MI of each residue for the WT and A) the active variants as well as B) the inactive variants. The WT is represented as a black cross and each variant is represented by a unique color, 12 colors for the active and 56 colors for the inactive.....	101
2.2	Figure 4.6 Comparison of the total MI for the WT and A) the active variants as well as B) the inactive variants. Residues where the total MI of a variant's residue is found to be more often greater than that of the WT are shown as red. Residues where the total MI of a variant's residue is found to be more often less than that of the WT are shown as blue. Residues where the total MI of a variant's residue is found to be more often similar to that of the WT are shown as green	103
2.2	Figure 4.7 Confusion matrix of the Logistic Regression model performance A) the performance result of training set B) the performance result of test set with the recall, precision and F1-score shown at the bottom right.....	106

CHAPTER I

INTRODUCTION

Protein Fundamentals


Macromolecules, such as proteins, are dynamic structures made from polymerized amino acid compounds. Just like electrons, neutrons and protons are the building blocks of an atom/element, amino acids are the building blocks of a protein. Amino acids are organic compounds, with the 20 most common amino acids being composed of 5 primary elements (carbon, nitrogen, oxygen, sulfur, and hydrogen)⁶⁻⁸. Each amino acid has an amino group and carboxyl group bridged by an alpha carbon ($C\alpha$), Figure 1.1A. The most distinguishing characteristic of an amino acid is the side chain (R group) stemming from the $C\alpha$; each side chain has unique chemical and physical properties that make a protein capable of adapting to their environment and purpose of completing a necessary function⁶⁻⁸, Figure 1.1A. Each side chain consists of a different number and/or arrangement of atoms with a terminal functional group such as hydroxyl, carboxyl, and alkyl group as well as a few special cases⁶⁻⁸. These functional groups generally provide the protein the ability to function as both hydrophilic and lipophilic mediators, meaning they are well suited to function in a polar solvent (water) as well as interact with non-polar substrates or lipids⁶⁻⁸. An amino acid sequence is translated from tRNA (aminoacyl-transfer ribonucleic acid) and mRNA (messenger ribonucleic acid) through the process of a condensation reaction, where the

A)



C)

D)



2

Figure 1.1: Shows the **A)** basic amino acid structure, **B)** amino acid chain formation, **C)** dihedral angle, and **D)** secondary structure elements.

The next step for a polypeptide is the formation of secondary structure elements, these are unique characteristics of the backbone that form due to the rotation about the plane (highlighted regions) that contains the peptide bond, Figure 1.1B. These rotations are measured as the dihedral angle or the measure of rotation (ϕ and ψ angle) about the peptide bond between neighboring alpha carbons¹¹⁻¹⁴, Figure 1.1C. Essentially, the lowest energy state is when both Ca are 180° apart (anti) which is referred to as a trans conformation, however the inclusion of the side chain and steric interactions produces various angles that can be more favorable as residue interactions occur¹¹⁻¹⁴, Figure 1.1B&C. There are 4 major secondary structure elements (α -helices, β -sheets, coils and turns) with the α -helices and β -sheets being more rigid in nature as they tend to be more ordered, have less mobility and share in hydrogen bonding networks¹¹⁻¹⁴, Figure 1.1D. Coils and turns are disordered sections of polypeptide chain that tend to have a greater range of mobility due to less restriction of rotation of the ϕ and ψ angles^{15, 16}, Figure 1.1B-D.

The collection of secondary structure elements is considered as a tertiary structure or the folded polypeptide chain contrived of a complex hydrogen bonding networks and dynamic residue interactions¹⁷⁻²⁰, Figure 1.1D. Some tertiary structures can function independent of additional configurations and are termed as a monomeric chain^{18, 21-23}, Figure 1.1D. However, most monomers can and will interact with one or more other monomers through oligomerization via non-covalent protein-protein interactions^{3, 18, 24}.

The protein databank (PDB)^{6, 10, 25, 26} reports that the majority of the known proteins are asymmetrical, which is weighed heavily on monomers though dimers or higher-level oligomers of the structure may exist. Consequently, oligomeric proteins can form via homomeric (2 or more of identical amino acid chains) or heteromeric (2 or more non-identical amino acid chains) interactions with other monomers^{6, 10, 25-27}. Homomeric proteins are found to have symmetry whereas heteromeric proteins are asymmetric, with homomeric proteins making up the majority of the observed proteins in a cell ^{6, 10, 25-27}. It is highly efficient for multiple homomeric subunits to come together and operate as a collective entity, particularly in the case of structural components of the cell ^{6, 10, 25-27}.

A protein is a fundamental component of biological systems that make up the structural component as well as the hardware and machinery that respond to stimuli ^{6, 10, 25-27}. The role of a protein is encoded based on its amino acid sequence, evolved such that, minor changes to the amino sequence can alter the output of the system and drive the system to be unstable ²⁸⁻³⁰. Proteins can be classified into several groups that are known to perform specific processes within a biological system ^{6, 10, 25-27}. Structural proteins are fibrous proteins that make up the skeleton of the cell. While transporters, antibodies, messenger, and receptor, as well as enzymes are a major part of most metabolomic pathways for a biological system ^{31, 32}. In these studies, I focus on enzymes. Enzymes are a specific type of protein known for carrying out necessary catalytic processes on small and large bio/molecules; depending on their purpose they can either promote or disrupt cellular function ³³⁻³⁷. Enzymes are highly important and efficient at transforming a

substrate into a necessary product ³³⁻³⁷. There are a number of different types of enzymes, but the research presented here focuses on a protease ³³⁻³⁷. A protease is responsible for breaking down other proteins, a particular important function that can be used for the removal or conversion of un/necessary proteins ³³⁻³⁷.

The most common amino acid sequence that makes up a protein structure is termed the wild-type (WT) sequence. The low frequency side chain fluctuations from each residue of the WT are uniquely characteristic of that amino acid sequence ^{38, 39}. Thus, the WT is encoded and tuned to perform a specific task to meet the demand of the system. However, due to improper translation/transcription process substitutions, insertions or deletions of an amino acid occur producing variant structures ⁴⁰⁻⁴³. Variant structures retain most of the amino acid sequence from the WT depending on the number of edits to the amino acid sequence that occur but for simplicity we will focus on single point substitutions. A change from a single point substitution may account for about a 1% difference from the WT, but can result in variants that can range from having greater or no activity towards a common substrate ⁴⁰⁻⁴³. The consequences that these conformational changes have locally that affect another non-adjointing region can be addressed by characterizing the alterations caused by packing defects. Changes caused by a single point substitution are often not directly connected to a known active site and some are deemed as allosteric “hotspots” ⁴⁴⁻⁵⁰. Allostery is a process where a change that occurs at one site influences the activity of another non-adjointing site ⁴⁴⁻⁵⁰. The mechanism of action for allosteric regulation is not very well understood, especially for changes that occur at distant sites. The study of intra-

protein communication, networks, regulation, and stability can provide scientists the methods to design and understand drug delivery, disease, and cellular processes better⁵⁰⁻⁵⁴. Allosteric processes are found to be highly influential towards viral and disease progression as well as regulating the activity of an enzyme in order to preserve or disrupt cellular function⁵⁰⁻⁵⁴. There are a number of approaches developed towards interpreting how allosteric affects propagate, though replication of the results or identification in other novel systems is often difficult.

When performing in-silico modeling of protein structures the Protein Data Bank (PDB)^{25, 26, 55-62} is one of the largest repositories of data regarding protein structure details. The structural details for a protein are obtained through analytical experimental methods such as nuclear magnetic resonance (NMR) and single-crystal X-ray diffraction methods that are capable of determining the position and arrangement of the elements of a structure with a resolutions close to 3.5Å or better^{25, 26, 55-62}. PDB files contain the structural details that account for the 3-dimensional position of each element in the structure^{25, 26, 55-62}. Most of the experimental processes used to identify a protein structure use either cryogenic temperatures or other processes that can trap potential energy in the crystalized protein structure. Therefore, when inputting a PDB file into an in-silico model it is essential to take precautionary measures to minimize, relax and temper the excess stress trapped in the crystal structure to ensure a more desirable outcome from the statistical methods⁶³⁻⁶⁹. After minimizing and tempering the system, the protein can now be studied using molecular dynamics or other statistical methods to identify the mechanism of interest.

Molecular Dynamics

The use of well-developed theoretical models combined with evolving computational methods offers computational scientists the capabilities to advance the knowledge of the dynamic characteristics of protein structures⁷⁰⁻⁷². Numerical techniques, such as molecular dynamics (MD) simulations, grant theorists the ability to investigate changes in the dynamics of a system to acquire strong clues towards mechanisms that influence the function of a system^{52, 73-78}. The studies represented here focus on conducting atomistic molecular dynamic (MD) simulations using NAMD, where all the atoms of the system are accounted for to achieve the highest amount of detail of the system^{79, 80}. MD simulations can be performed using either an atomistic (all atom) or coarse-grained (reduced representation) model, coarse graining is different as it uses a reduced representation of the system^{79, 80}. An atomistic model is best suited for capturing refined detail of a system but is computational more costly, whereas a coarse-grained model offers less refined detail and is often used for identifying general characteristics that require less computational cost⁸¹⁻⁸³. Both models implement conditional parameters intended to approximate the environmental conditions to acquire a reasonable outcome (dynamics) of the system⁸¹⁻⁸³. A thorough analysis of the MD production must be performed in order to support whether the system has converged or equilibrated^{47, 54, 84-86}. Using a well equilibrated system can strengthen the accuracy and certainty that the statistical analysis techniques applied offer a legitimate assessment of the system^{47, 54, 84-86}. In summary, the

use of a well composed in-silico model can be used to better understand certain mechanisms of action, like allosteric communication, impact the function of an enzyme.

There are various methods developed that can be used to analyze the production from a MD simulation with the intent of extracting, isolating, and uncovering beneficial details of the components of a system ⁶³⁻⁶⁹. The output of an MD simulation combined with the chemical and physical properties associated with the system is highly dimensional and requires well thought out dimensionality reduction techniques in order to identify trends and make the output more comprehensible. The first approach taken looks to characterize the mobility of the system by seeking to identify the allosteric pathways along with allosteric “hotspots”. Mobility is a multifaceted component of a protein such that the global structure requires flexibility and rigidity to adjust to the changes when interacting with a substrate as well as the solvent ⁸⁷⁻⁸⁹.

Contacts

A substitution alters the space, voids, or cavities that exists in the WT creating a reduction or expansion to the surrounding area that can cause denser or looser packing throughout the variant ⁹⁰⁻⁹⁷. The realignment caused by a substitution can change the residue contacts that are common to the WT as a result from the adjustments to the packing defects⁹⁰⁻⁹⁷. The adjustments to the internal packing can alter the interface between the protein and the solvent, the interface between small/large molecules, or respiratory actions

observed from the pulsation of the low frequency normal modes at longer timescales⁹⁰⁻⁹⁴. Contact maps from the side chain interactions of the residues within a protein are commonly used by researchers as a means to interpret significant changes caused by realignments of a substitution⁹⁰⁻⁹⁷. From a theoretical standpoint, a contact is seen to occur within a structure when 2 residues come within $\sim 3\text{\AA}$ of each other, the median distance found for hydrogen bonding to occur⁹⁰⁻⁹⁷. Similarly, van der Waals interactions, though weak, dictate as two objects come closer together, due to non-attractive forces, the more they are repelled apart^{34, 98-100}. In other words, if more residues are coming in contact with one another the mobility of the structure is constrained. Lowered van der Waals contact densities have been connected to less thermostability and accounts for a looser packing density of viral proteins, especially for RNA viral proteins^{95-97, 101}. Consequently, if an enzyme gains or loses contacts it should adversely affect the function of an enzyme.

Many of the current approaches employed to investigate the mobility along with the packing defects question the thermodynamic stability of the structure, particularly for extremophilic proteins^{88, 102-104}. Extremophilic proteins, such as thermophilic or barophilic proteins, are adeptly unique as they can operate at higher temperatures and pressures than that commonly observed in mesophilic organisms^{88, 102-104}. Mesophilic proteins thrive at moderate temperatures and pressures that support terrestrial life^{88, 102-104}. Both thermophilic and mesophilic proteins have a hydrophobic core, but it is suggested that hydrophobic properties of the core found from thermophilic proteins allow for much more compacted core^{88, 102-104}. Having a more compact core would suggest that a thermophilic protein loses

the nature of packing that supports the function of the WT under mesophilic or moderate temperatures and therefore show an overall increase in contacts^{88, 102-104}.

Stability

The change in Gibbs free energy (ΔG or dG) is able to characterize the stability of a protein and commonly approached as the root for explaining the mobility of the system^{33, 105-107}. MD simulations can be used for the study of how mechanical and thermal energy is able to perform work by a system^{33, 105-107}. Therefore, evaluating the change in Gibbs free energy (ΔG) contains components regarding the mobility of the structure, but also includes details relating to other features^{102, 103, 108-110}. By resolving the binding or folding free energy change upon substitution ($\Delta\Delta G$ or ddG) researchers offer insight into the influences of a single point substitution^{102, 103, 108-110}. It is observed that the electrostatic interactions between the binding pocket and the substrate are often the primary source acting on the binding of a substrate to an enzyme, though the dynamics plays a contributing role^{102, 103, 108-110}. Research suggests that the changes in stability caused by a single point substitution supports the observed free energy change upon substitution ($\Delta\Delta G$)^{102, 103, 108-110}. However, what is actually observable doesn't adequately resemble the thermodynamic profile of a protein structure in-vitro/in-vivo^{102, 103, 108-110}. Thus, providing a reasonable characterization of the thermodynamic stability of a protein is regarded as a highly sought-after concept by most theoretical bio-physicists and bio-chemists. Can the change in stability offer insight towards the function of a protein and its variants? Does stability offer

any detail towards the rigidity or flexibility of the side chain motion? Consequently, the correlations between protein function, stability and mobility need to be addressed adequately.

The mechanistic nature of thermodynamic stability is very complex. Too often it is thought of as absolute conditions, where a protein folds or unfolds, binds or does not bind^{34, 89, 111-113}. It can be quite certain a protein will fold or unfold when faced against extreme environmental conditions, but of course binding and catalysis is a much different aspect^{34, 89, 111-113}. Mostly, extreme environmental conditions can be thought of as higher or lower than normal temperatures as well as higher or lower than normal pressure. These extreme environmental conditions can be unfavorable for proteins and can lead to thermo-denaturing. While there are other situations, like chemical denaturing, that can drive a protein to become unstable or even unfold^{34, 89, 111-113}. Some variants lack the functional characteristics of an active counterpart (WT or CA variant) and the question often asked is whether or not it is able to fold. Non-functional variants are complex and though they lack activity towards a common substrate they may have preference towards different target substrates^{34, 89, 111-113}. In some other cases non-functional variants may be unable to fold but they may also be in a fixed opened or closed conformation that is unsuitable for normal function^{34, 89, 111-113}. When considering the completion of all of the steps for the binding of substrate followed by catalysis then release of a substrate, it can only occur from an active variant (including the WT), Figure 1.2. It is very unlikely for an inactive variant to complete all 3 steps for the catalysis of a substrate, considering the completion of all three steps is

reserved for the WT and active variants. It is more likely that an inactive variant only completes one or maybe two of the three main steps (please see Figure 1.2 and the corresponding caption) for the catalysis of a substrate^{34, 89, 111-113}. An inactive variant may be able to bind and release (but not catalyze) or it can bind and catalyze (but not release) as well as be completely non-interactive with a substrate^{34, 89, 111-113}.

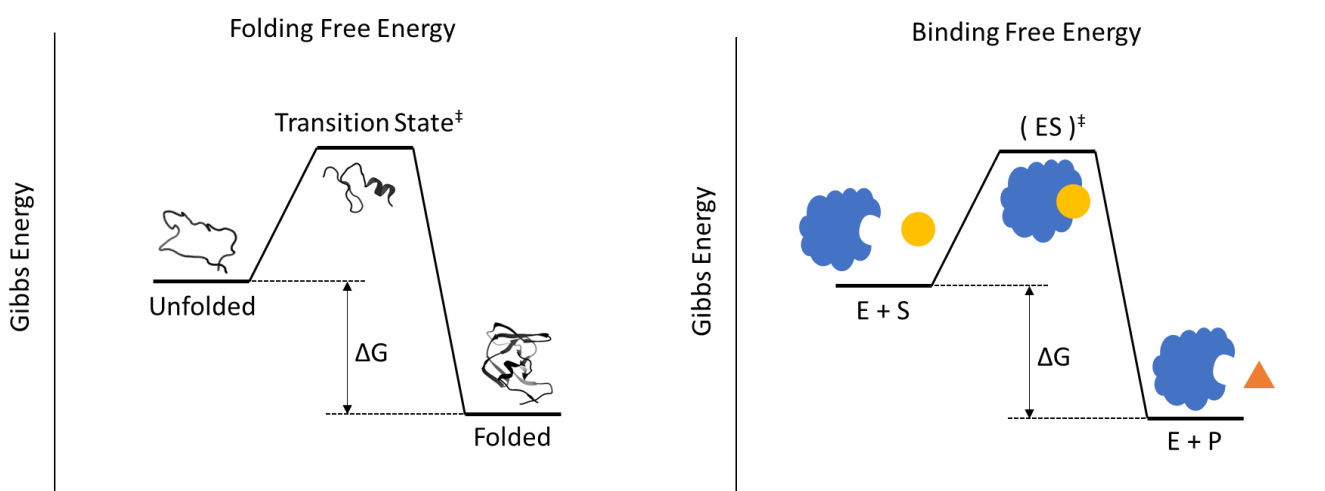


Figure 1.2: Reaction diagram for the folding free energy change (A) and binding free energy change (B) of a protein structure. Figure 1.2 (A) shows the three steps of folding of a protein; (1) unfolded monomer, (2) the formation of secondary structure elements, and (3) the completely folded monomer. Figure 1.2 (B) shows the three steps associated with the binding and catalysis of a substrate; enzyme (E) + substrate (S), the ES complex, and the E + product (P).

In general, it would be more logical to characterize the free energy change of binding or folding upon substitution ($\Delta\Delta G$) as a scale of mobility, how flexible or rigid is the structure^{34, 89, 111-113}. Essentially, if the folding/binding free energy is found to be stable or unstable it cannot be considered as conclusive, such that it will or will not fold or bind the substrate. In many ways, much of the literature and resources used to evaluate the folding or binding free energy of a protein include additional methods that characterize the

flexibility of a structure, but there are still many challenges and variables that need to be addressed to best characterize a protein's mobility^{109, 114-116}.

Dynamic Cross Correlation

The next method approached to examine the MD runs and structure of HIV-1 protease is a dynamic cross correlation (DCC)¹¹⁷⁻¹²⁰. DCC is a method that can be used to identify the pairwise linear dependency of two components from a single conformational state or the ensemble of conformational states from the dynamics of a protein^{65, 121-125}. The DCC pairwise linear dependency function normalizes the relationship between two components¹¹⁷⁻¹²⁰. If component 1 is at position i and component 2 is at position j , the normalization is described by the Pearson correlation coefficient between position i and j ranging from 1 (correlated) to -1 (anti-correlated). The output from a DCC is a r^2 symmetrical matrix, where r is the number of residues¹¹⁷⁻¹²⁰. A major benefit from a DCC matrix is that rigid secondary structure elements tend to be highly correlated and easily show disfunction¹¹⁷⁻¹²⁰. Consequently, there are r^2 data attained from a DCC along with the fact that MD simulations of the same system can vary slightly each time it is conducted identification of significant changes versus systematic noise makes it tenuous to directly compare a larger number of variants^{47, 126-130}. Many researchers have proposed the evaluation of linear correlations could be a possible means to interpret catalytic function and/or observation of the allosteric pathways within a protein^{47, 126-130}.

Mutual Information

An alternative method to DCC for investigating the residue interactions within a protein is Mutual Information (MI), which allows for the identification of both linear and non-linear mutual dependency between two variables¹³¹⁻¹³³. MI is a statistical method within information theory that focuses on the communication between data sets^{131, 132, 134-136}. The theory behind MI was developed by Shannon in 1948; Shannon defined entropy as a measure of information, choice and uncertainty, associated with the variables being studied, analogous to the concept of entropy from thermodynamics^{131, 132}. Many have adopted this theory to understand the uncertainty and noise associated with a biological, chemical, or physical phenomena^{67, 134-137}. It is remarked that as the choice becomes uncertain the entropy increases, greater than 0, information contained within an outcome increases as the uncertainty of the system/observation increases due to more possible outcomes^{131, 132}, eq. 1.1. When considering joint entropy ($H(R_i;R_j)$), the uncertainty remains the same as Shannon entropy, whereas knowing more about the shared states of residues i and j decreases the joint distribution, eq. 1.2^{67, 134-137}.

$$H(R_i; R_j) = - \sum_{(r_i \in R_i)} \sum_{(r_j \in R_j)} p(r_i, r_j) \log p(r_i, r_j) \quad \text{eq 1.1}$$

$$I(R_i; R_j) = \sum_{(r_i \in R_i)} \sum_{(r_j \in R_j)} p(r_i, r_j) \log \frac{p(r_i, r_j)}{p(r_i)p(r_j)} \quad \text{eq 1.2}$$

Shannon entropy can be applied to acquire details that can be used to identify the coevolutionary dependency in multiple sequence alignments of residue pairs, something that was indistinguishable from the previous study of DCC^{67, 134-137}. Most co-evolution studies seek to identify residues within a specific distance, often neighboring, that coevolve^{67, 134-137}. Some of the co-evolution studies question whether distant residue pairs coevolve as well but not much has been found to support this theory^{67, 134-137}. Much of the information discussed in the literature speculates it can be possible, but they also postulate that distant residue communication is less of a signal and mostly background noise¹³⁸⁻¹⁴⁰.

Consequently, if residue *i* affects residue *j* an allosteric effect may not be evident and a deeper understanding of the topological features that effect distant residues and ultimately a protein's function should be further evaluated. Residues can pack close together and provide a means for the amino acid sequence to generate the secondary structures²⁸⁻³⁰. Each residue has specific chemical features that effect the surrounding environment and since proteins are dynamic, the degrees of freedom between residues can vary depending on the neighboring residues²⁸⁻³⁰. Within a protein structure each amino acid residue experiences many different environmental conditions and interactions, such that surface residues mostly interact with the solvent, bulk and interface residues primarily associate with other residues, and binding pocket residues come in contact with the solute or solvent, if accessible^{93, 141-145}. Consequently, the forces and interactions imposed on each residue can vary significantly, depending on the location and the neighboring

residues, constituting a considerable factor for the identity and/or pathway of allosteric communication^{23, 50, 76, 146}.

HIV-1 Protease

To better understand how the function/activity of an enzyme, human immunodeficiency virus type 1 (HIV-1)^{1, 34, 147-154} protease, and its variants can be modified from allosteric interactions we analyze the dynamic changes of the protein and variants with unique single point substitutions¹. Incidentally, HIV-1 protease is a promiscuous enzyme and the experimental data¹ evaluated here only characterized HIV-1 protease with one of its potential target substrates. Using well-defined statistical methods and techniques we seek to develop a clear understanding of how allosteric interactions are processed from one residue to the next within the symmetrical robust HIV-1 protease^{1, 34, 147-154}.

In my study, the structures of the HIV-1 protease variants with a known experimentally reported¹³ activity are exhaustively tested to better understand how dynamics influences allosteric signaling. Probing the dynamic interactions of the residues of an enzyme and its variants can offer insight of the developmental importance for allosteric signaling and their connection to a protein's function^{44, 74, 97, 155-165}. The chapters in this study were designed as following: Chapter II we investigate how a single point substitution alters the conformational landscape of HIV-1 protease variants^{13, 109, 165-175}. The conformational changes obtained from MD simulations were evaluated using RMSF,

the residue contacts between non-neighboring residues and the folding/binding free energy changes upon a corresponding substitution to explore the relationship between the mobility of the structure upon a single point substitution and the protein's function. Chapter III, we seek to understand whether the residue positions found to adopt significant correlations have any evolutionary importance based on their relationship towards the function of the HIV-1 protease variants^{13, 109, 165-175}. We examine the vitality of these evolutionarily important residues using dynamic cross-correlation (DCC) to identify if there is a relationship between experimental information¹ (variant activities) and the protein dynamics of HIV-1 protease. The identification of any significant constraint imposed on the dynamics can lead researchers to expose potential allosteric sites found to disrupt the catalytic activity of the variant. Recognition of any substantial association between the dynamics of specific residues and allosteric communication or mechanism requires detailed examination of the dynamics of HIV-1 protease and its variants^{109, 165-175}. In Chapter IV we explore the non-linear dependency using the Mutual Information (MI) between each residue pair of the variant structures of HIV-1 protease, to investigate how the orthogonal information can assist us to understand the change in protein function upon single point substitution. We also consider how MI can be used as a metric for identifying allosteric "hotspots". Due to the robust nature of HIV-1 protease and the dimensionality of the results, we are also able to predict the protein function/activity change caused by a single point substitution given the MI from the MD simulation of the protein using ML algorithms. Effectively the mobility, dynamics, and non-linear features tested in these

studies are found to be useful towards the prediction of potentially drug resistant substitutions related to the catalytic efficiency of HIV-1 protease and the variants.

CHAPTER II

CHANGES IN THE MOBILITY AND DYNAMICS WHICH IMPOSE CONSTRAINTS ON THE FUNCTION OF A PROTEIN

Introduction

Macromolecules, such as proteins, are dynamic structures that operate as the fundamental machinery of biological systems needed to ensure optimal biological function of a host ²⁸⁻³⁰. Proteins are versatile biological tools used to form, align, transcribe, as well as catalyze cellular/metabolic components in support of proper cellular function of the host ^{31, 32}. Here we focus on enzymes, particularly enzymes found to regulate the viral lifecycle of HIV by catalyzing the matrix capsid (MA/CA) polypeptide. Enzymes are particularly useful at either promoting or disrupting cellular function, and in many ways, it is an essential part of their job ^{31, 32}.

Proteins are composed of amino acids linked together by peptide bonds forming a polypeptide chain ²⁸⁻³⁰. Each amino acid residue is joined from an N (amine end) to C

(carboxyl end) terminus with a side chain stemming from the C α (alpha carbon)²⁸⁻³⁰. The C α is the central link from the side chains that makes up the backbone of the polypeptide chain. There are many distinguishing chemical and physical characteristics of an amino acid, with the functional groups associated with the side chain dictating their general nature. Each amino acid residue provides a unique local environment derived from the chemical and physical properties of the functional groups associated with its side chain^{31, 32}. The chemical and physical features of any residue contribute to the thermodynamic nature of the local and global environment²⁸⁻³⁰. A protein takes advantage of each residues local environment that are, in turn, adapted to govern a protein's global behavior^{31, 32}.

The most common sequence of amino acids for a protein is known as the wild-type (WT) structure. Though during the transcription process alterations to the amino acid sequence is often changed that leads to the rise of variants of the protein due to substitutions, deletions, or insertions. Presently we focus on single point substitutions as they are the common case study towards assessing the evolutionary function of a protein^{54, 159, 166}. Variants can be found to express greater, less, or no catalytic function towards a common substrate, with regard to the activity expressed by the WT^{54, 159, 166}. A single point substitution can transform the function of local environment and can also influence changes to occur at another non-adjointing distant site/s of the enzyme, which is known as allostery⁴⁴⁻⁵⁰. The study of single point substitutions can aid in the classification of regulatory pathways and networks that dictate the role of a protein's dynamics^{109, 167-169}. Understanding these regulatory processes can outline how the gain or loss of function can

be correlated with a protein's dynamics^{109, 167-169}. Interpreting how allosteric effects propagate, why a change at "A" can affect the function at "Z", is a major area of focus within the bio-physical sciences⁴⁴⁻⁵⁰. In general, a single point substitution only makes up 1% or a smaller fraction of the protein structure, yet the influences on the functionality of the global structure results in variants with greater or no activity^{54, 159, 166}. Alternatively, much of the relative interactions from the common residues between the variant (VT) and WT continue to stimulate normal modes of activity and can suppress any observed allosteric behavior^{54, 159, 166}. Nonetheless, substitutions are highly important for regulating the activity of an enzyme as a means to preserve cellular function⁵⁰⁻⁵⁴.

Substitutions allow for a local realignment/adjustment that effects the adjoining neighboring residues and secondary structure, these local realignments can also cause changes to occur at a distant residues or regions of the structure^{90, 142, 170-172}. These adjustments disrupt the stability of a structure by promoting rigidity or flexibility at the substitution site as well as other regions of the protein^{90, 142, 170-172}. Some residues can be characterized by their ability to promote such changes, such as glycine, being the smallest amino acid, is more often found to increase flexibility. Whereas larger amino acids can be considered to cause more rigidity, however the impact any one residue has is not very well understood. This is mainly due to the nature of the exchange, substitutions that alter the function of the enzyme vary depending on the chemical and physical nature of the residue exchange^{42, 173-178}. Significant efforts have been made to examine a protein's dynamics in order to interpret how a residue can be linked to an allosteric "hotspot"^{107, 179-181}.

Interpreting how these hotspots work can assist in enhanced targeting of druggable sites by either enabling or disabling an enzyme's function allowing for improved control and treatment over a disease ^{107, 179-181}.

Here we study the human immunodeficiency virus type 1 (HIV-1) protease and several of its variants to better understand how allosteric hotspots influence a protein's dynamics. HIV-1 protease is an essential enzyme in the maturation process for the viral replication in the immature AIDS (acquired immunodeficiency virus) virus ¹⁸². HIV-1 protease is responsible for the cleavage of the matrix-capsid (MA/CA) substrate and eleven other sites from Gag and Gag-Pol poly protein complexes required for the viral replication process¹⁸². HIV-1 protease has two identical 99-amino acid residue chains that make up the robust C2 symmetrical dimer ¹⁸³⁻¹⁸⁵. Symmetry has been observed to contribute to the evolutionary stability and cooperative interactions between subunits of many oligomeric proteins, like that of HIV-1 protease ^{54, 186-190}. HIV-1 protease is scrutinized heavily in both theoretical and experimental trials, particularly that of drug trials of protease inhibitors, due to its robustness and role in the viral life cycle ¹⁹¹⁻¹⁹³. HIV-1 protease can rapidly produce functional and nonfunctional variants due to the inefficient replication process that allows for competing variant structures ¹⁹¹⁻¹⁹³. Some of these competing variant structures can lead to drug resistance and poor binding of protease inhibitors ¹⁹¹⁻¹⁹³. In some cases, the competitive variants can be preferentially selected for viral replication in the presence of a protease inhibitor drug ¹⁹¹⁻¹⁹³. This preferential selection can be thought of as enzymatic evolution.

Here we institute molecular dynamic (MD) simulations to capture the low and high frequency harmonic and anharmonic motion from the side-chain fluctuations of an enzyme, HIV-1 protease^{122, 194-198}. Experimental data is a greater resource for interpreting how allosteric modulations can be used to regulate the activity of competitive variants of an enzyme like HIV-1 protease. Martinez et. al.¹ examined the catalytic activity of 107 variants of HIV-1 protease towards the cleavage of matrix-capsid (MA/CA) polypeptide domain. The variants are found to have a single amino acid substitution repeated in each chain of the C2 symmetrical dimer, Figure 2.1.

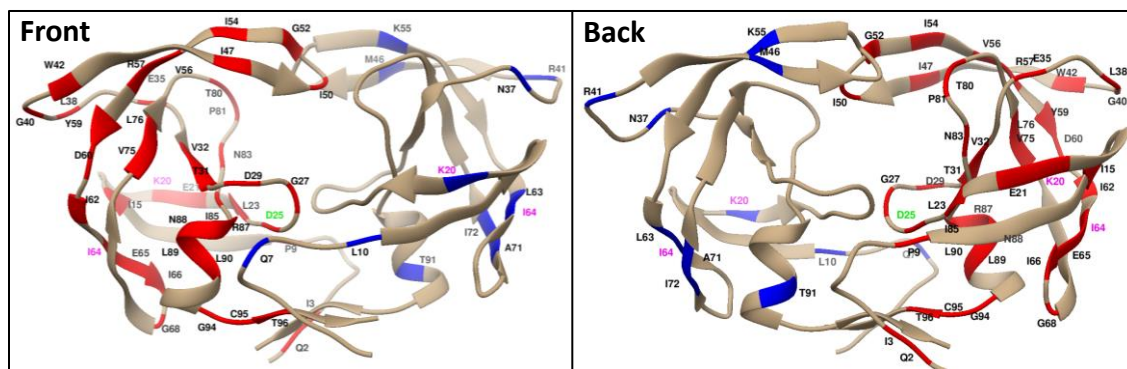


Figure 2.1: The figure shown above is a front and back view of the HIV-1 protease (PDB id: 1hxw). The colored mutant residue positions are highlighted in a single chain, though each substitution was applied to both chains, to show the location of the catalytically inactive (CI) and catalytically active (CA) variant residue positions (red and blue respectively). Of the 99 common residues that make up each chain, there are 42 unique positions are found to cause an inactive variant and 12 positions are found to cause activity equal or greater than the WT structure¹. There are two residue positions that have both zero and greater activity, residues K20 and I64. The K20 CA residue is K20R, a basic-to-basic side chain substitution, but the K20 CI residue is K20E, a basic to acidic side chain substitution. Similarly, for I64 CA residue is I64V, a non-polar to non-polar side chain substitution, whereas the CI residue is I64T, a non-polar to polar side chain substitution. There are also a number of redundant positions that are shown to have zero activity¹.

The in-vitro experimental data¹ presents variants that can be classified as either having equal to or greater than, moderate, or have zero activity compared to the WT. Here we focus on the 12 variants characterized as having equivalent or greater activity (colored blue in Figure 2.1) and the 56 (42 unique residues) variants characterized as having no observable activity (colored red in Figure 2.1) compared to the WT structure, Figure 2.1. Substitutions like these are evolutionarily important for regulating the activity of the viral protease. Since enzymes are very good at catalysis, the presence of non-functional variants ensures that the functional variants can still operate in the presence of drug inhibitors.

Molecular dynamics makes it possible to characterize how mechanical and thermal energy from the bonds and atoms of a system is transformed into work^{102, 103, 108-110}. In the case of a protein the most predominate source of mechanical energy comes from the conformational changes from the amino acid side chains that result in fluctuations of the backbone of the structure^{102, 103, 108-110}. Our research concentrates on the transfer of mechanic and thermal energy to/from the local and global environment using uniform environmental conditions meant to simulate the standard conditions that promote the regulatory function of amino acid substitutions^{102, 103, 108-110}. Much of the current literature utilizes the nature of thermophilic proteins through applied temperature gradients to characterize changes in nature of a protein^{89, 199, 200}. Thermophilic proteins are adeptly

unique as they can operate at higher temperatures, not commonly observed in normal cellular conditions of mesophilic proteins^{88, 102-104}. It is also suggested that in many instances, the thermodynamic properties of thermophiles are not directly comparable to mesophilic proteins, such as HIV-1 protease^{88, 102-104}.

The amount of thermodynamic energy of a system is commonly measured using the change in Gibbs free energy (ΔG or dG) and can be used to understand its thermostability^{102, 103, 108-110}. In order to understand the influence of a single point substitution, researchers seek to resolve the binding or folding free energy change upon substitution ($\Delta\Delta G$ or ddG)^{102, 103, 108-110}. The binding free energy of an enzyme with a substrate is often found to be associated with the electrostatic interactions between the binding pocket and the substrate, though the dynamics plays a contributing role^{102, 103, 108-110}. However, the changes in free energy change upon substitution (ddG) can provide evidence to support the change in stability that is often observed to influence the activity of an enzyme^{102, 103, 108-110}. Being able to equate the thermodynamic stability of a protein structure to what is actually experimentally observable is highly sought-after goal for most theoretical bio-physicists and bio-chemists. It is common to associate the thermodynamic stability of a protein with whether it will fold or unfold as well as function, but it is more complex than that. If a decrease in the free energy change upon substitution (ddG) is found to be stabilizing can this provide clues towards the function of a variant? Does stability offer any detail towards the rigidity or flexibility of the side chain motion? Consequently,

the correlations between protein function, stability and mobility need to be addressed adequately.

It is apparent that the residues that make up the WT structure are essential for ensuring proper function of the protein, such that changes in any one residue can alter both the local and global environment^{49, 50, 113}. Being able to identify how these modifications can influence the stability of the structure can be used to uncover the regulatory nature of hotspots (key residues) and offer insight towards the nature of viral proteins, like HIV-1 protease. Quite often the free energy change of binding or folding upon substitution (ddG) is used to interpret whether a residue substitution will impact the function of a protein^{23, 201-203}. Consequently, though much of the resources available can provide a reasonable prediction of the free energy change of binding or folding upon substitution (ddG) they are still highly disputed^{23, 201-203}.

Thermodynamic stability is too often thought of as finite scenarios, where a protein folds or unfolds, binds or not but the mechanistic nature is very complex^{34, 89, 111-113}. When including the extrema of the environmental conditions it can be quite certain a protein will fold or unfold, but of course binding and catalysis is a much different aspect^{34, 89, 111-113}. There are a number of situations that can drive a protein to unfold, and some select scenarios where it can't fold even in adequate environmental conditions^{34, 89, 111-113}. Under normal conditions a variant should fold but may lack the functional characteristics of an

active counterpart (WT or CA variant) or have a different target substrate^{34, 89, 111-113}. In some other cases, the non-functional variants may be in a fixed opened or closed conformation that are unsuitable for normal function^{34, 89, 111-113}. When considering the binding of substrate then catalysis followed by the release of a substrate, it can only occur from an active variant (including the WT), but it is not impossible for an inactive variant to bind a substrate plus one and only one of the other subsequent steps^{34, 89, 111-113}. The challenge comes from having the third step, an inactive variant may be able to bind and release (but not catalyze) or it can bind and catalyze (but not release)^{34, 89, 111-113}. In general, the free energy change of binding or folding upon substitution ($\Delta\Delta G$) cannot be considered as finite, if the fold/binding free energy is stable or unstable, it would be more logically to characterize it with a scale of mobility, how flexible or rigid is the structure^{34, 89, 111-113}. In many ways much of the literature and resources used to evaluate the folding or binding of a protein include the characterization of the structures flexibility, but there are still many challenges and variables that need to be addressed to best characterize a proteins mobility^{109, 114-116}.

Here we propose that an increase or decrease in the mobility of the residue side chains can be found to be adversely related to the activity of an enzyme. The root mean square fluctuation (RMSF) of each residue of a protein is a widely used method for determining the mobility within a protein²⁻⁵. RMSF for dimers, like HIV-1 protease, closely follow the trends observed from the temperature B-factor used in characterization of the X-ray crystal structure of a PDB file²⁻⁵. Furthermore, mobility can be modulated by

the realignment of secondary structure elements as well as the frequency of residue that come in contact with one another ²⁻⁵. Further, it can be postulated that if there are more contacts occurring within a structure the mobility is being constrained and therefore gaining novel contacts can negatively influence the function of a protein. Moreover, the change in free energy for the binding as well as the folding upon single point substitution for both the active and inactive variants are calculated to uncover if there is any relation between them to the function of the variants.

Methods

The following methodologies were used for modeling the WT and the variant structures of HIV-1 protease with the MA/CA substrate. Using the crystal structure from a PDB structure (PDB id: 1KJ4)²⁰⁸ that had the MA/CA substrate and a tetra-coordinated water molecule were extracted and placed into the binding pocket of the (PDB id: 1HXW)²⁰⁹ structure is observed to have the same amino acid sequence consistent with the WT structure ²¹⁰. Both PDB structures were aligned using superposition of the backbone atoms of the active site residues (within 4.5Å (angstrom) of the substrate) in order to place the substrate and water into the binding pocket of the WT structure. Using a software package called MODELLER ¹⁴. Each variant was created such that a single point substitution was generated in each chain of the appended WT structure. MODELLER¹⁴ applies a harmonic potential that is specific to the substitution for every bond, angle, and improper dihedral angle while all of the other residues for the protein are restrained ¹⁴.

The structures (WT, 12 active variants and 56 inactive variants) were solvated in a water box of TIP3P water molecules²¹¹, with the minimum distance of any atom of the protein structure to any face of the box 10 Å or greater. Potassium chloride (KCl) ions were included when the system was solvated to achieve a salt concentration of 150 mM (millimolar), keeping the total charge of the system equal to zero^{114, 212}. Any water molecules that have an oxygen atom closer than 2.6 Å from the protein and substrate were removed from the system^{114, 212}. The energy minimization and the MD simulations of the structures were performed using the molecular mechanics package CHARMM²¹³, with its 22nd version of the force field²¹⁴. A Particle-Mesh Ewald (PME) method was applied to treat the electrostatic interactions with a dielectric constant of 1 and a cutoff of 10 Å. The Lennard-Jones potential for any non-bonding interactions were set to 0 beyond 10 Å. Bonds involving hydrogen atoms were constrained using SHAKE²¹⁵ during energy minimization and MD simulation. The MD simulation was performed in an NPT (isothermal–isobaric; N stands for constant number of particles, P stands for constant pressure, T stands for constant temperature) ensemble, where the temperature was maintained to an average of 300 K (Kelvin) using a Langevin thermostat while the pressure was maintained to 1 atm (atmosphere) using the Langevin piston method²¹⁶.

The solvent of the system was first minimized for 4000 steps using a steepest decent (SD) algorithm while keeping the protein and substrate atoms constrained. In the following steps, a force constant of 1000 kcal/molÅ (kilocalorie/molecule) was applied as a distance-

based restraint to the substrate atoms and the atoms within 5 Å of the substrate from the catalytic site. The system was then minimized over 12000 steps using a SD algorithm while the distance-based restraint was scaled down to 800 kcal/molÅ. The solute atoms were then harmonically restrained with a force constant of 700 kcal/molÅ and minimized for 36000 steps using a SD algorithm while the forces constant was scaled down to 100 kcal/molÅ.

The system was gradually heated from 100 K to 300 K during 100 ps (picoseconds) of MD simulation with a 1 fs (femtosecond) time step with a distance-based restraint for the substrate and active site applied with a force constant of 150 kcal/molÅ. In the following steps, the atom-based restraint applied to the substrate and active site were removed to begin the MD simulation for the system. The MD simulation of the WT, 68 variants and 35 variants structures with moderate activity (greater than 0 and less than 95% of the WT activity) was run for 72 ns (nanosecond). Additionally, most MD simulations equilibrated after 4 ns making a total of 68 ns of production for each structure that was run in triplicate for a total of 204 ns of production.

Energy

To calculate the energy profile of each variant and the WT, chain A was isolated from the structure in order to calculate the folding free energy. Whereas in order to evaluate the binding free energy the structures were divided into the enzyme (E), enzyme-substrate (ES), and substrate (S). The molecular mechanics (MM) energy was obtained for each component (E, ES, S) using the output from the MD production run using CHARMM^{70, 72, 217, 218}. The maximum (max) and minimum (min) values of each energy term were taken from each MD production run in order to calculate each $d/(\Delta)$ energy term. For instance, the dMM (Δ MM) is calculated as the following:

$$dMM_{(ES, E, S, A)} = \max(MM) - \min(MM) \quad \text{eq. 2.1}$$

$$dMM = (dMM_{(ES)}) - (dMM_{(E)}) - (dMM_{(S)}) \quad \text{eq. 2.2}$$

Similarly, the Poisson–Boltzmann (PB) energy was calculated using delhipka²¹⁹⁻²²¹ for each component of the system.

$$dPB_{(ES, E, S, A)} = \max(PB) - \min(PB) \quad \text{eq. 2.3}$$

$$dPB = (dPB_{(ES)}) - (dPB_{(E)}) - (dPB_{(S)}) \quad \text{eq. 2.4}$$

The surface area (SA) energy term was calculated using VMD^{72, 222, 223}.

$$dSA_{(ES, E, S, A)} = \max(SA) - \min(SA) \quad \text{eq. 2.5}$$

$$dSA = (dSA_{(ES)}) - (dSA_{(E)}) - (dSA_{(S)}) \quad \text{eq. 2.6}$$

MMPBSA is used to calculate the change in free energy (dG) as well as the change in free energy upon substitution (ddG) using the following methods.

$$\text{MMPBSA} = \text{dMM} + \text{dPB} + \text{dSA} \quad \text{eq. 2.7}$$

$$\text{ddG} = \text{MMPBSA}_{(\text{Variant})} - \text{MMPBSA}_{(\text{WT})} \quad \text{eq. 2.8}$$

Contact maps were calculated using VMD^{72, 222, 223} by measuring the contacts or clashes that occur when any atom of residue i is within 3Å of residue j. Hydrogen bonding occurs between 2.7 – 3.3 Å, making 3 Å a reasonable cutoff as it is the median distance for hydrogen bonding to form based on a symmetrical distribution, though it can vary depending on the reactivity. However, we are looking for any form of contact occurring between the amino acid side chains, not just the potential of hydrogen bonding. Next the information was evaluated for significances by removing contact frequency that occur for less than 10% of the production length. Identification of common pairwise interactions was developed by eliminating redundant measurements where residue i (r_i) does not equal residue j (r_j) ($r_i \neq r_j$), while also removing neighboring residues ($r_{(i,j) \pm 1}$) for each variant case. A similar methodology was used to capture the common and novel contacts between each of the variants and the WT.

Chemical & Structural Properties of HIV-1 Protease

Table 2.1: Comparison of the Chemical and Physical Features for the WT and each of the 12 CA and 56 CI variants (y-axis) with the 99 amino acids of the HIV-1 protease (x-axis)

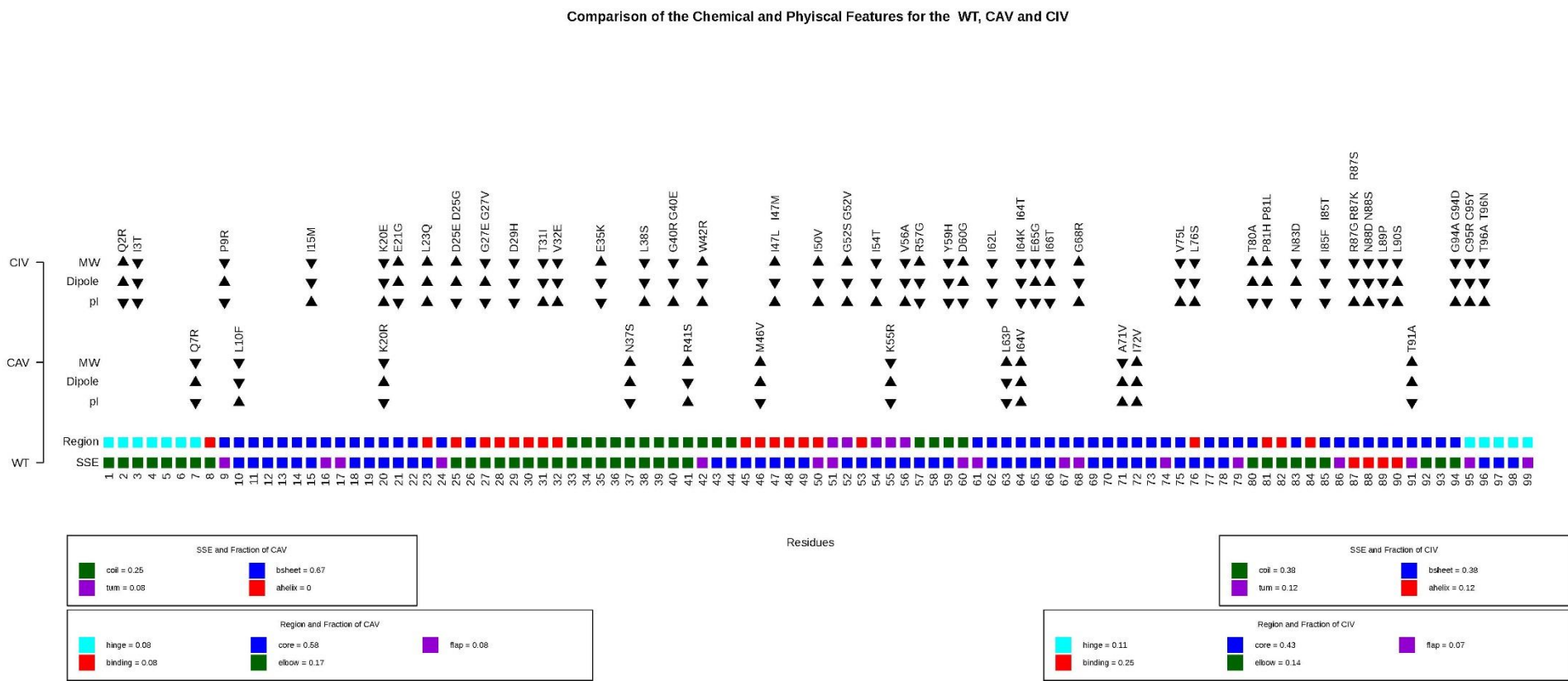


Table 2.1 seeks to outline a number of the relevant features related to the chemical and physical characteristics of the substitution for each variant. The legend on the bottom left and right are essentially the same, with the details regarding the CA variant are shown in the legend on the left and the CI variant are shown on the right. The color encoding is applied to the secondary structure element (SSE) that a residue is commonly found to belong. A second color encoding was generated for the region that the residue is observed to belong in relative to the global structure. Though we tend to think of these SSE as adding rigidity to a protein structure, they are dynamic in nature. These changes in the dynamic fluctuations can be seen with the alpha helix, chain A reports that residues 87-90 are responsible for forming the alpha helix but chain B reports that residues 87-92 are responsible for making up the alpha helix. Consequently, though a set of residues might be labeled as a being part of a specific SSE there is some uncertainty due to the dynamic nature of the residues.

There are several important substitutions to consider, that are substitutions from or to glycine and substitutions from isoleucine to leucine which are only observed within the CI variant cases. There are 6 substitutions to glycine and 8 substitutions from glycine making up 25% of the CI variants. There are 2 positions, I47L and I62L, where isoleucine and leucine have almost identical chemical features with the only difference being displacement of a methyl group from the gamma to the beta carbon, in either case resulting in inactive variant. Having the methyl group closer to the alpha carbon could help regulate

the mobility by limiting the conformational flexibility of the beta carbon. Table 2.1 dedicates a portion of detail towards the changes in molecular weight (MW), dipole moment, and the isoelectric point (pI) for each residue and the corresponding substitutions. An increase or decrease in any feature is represented by an upward or downward arrow, accordingly. About 60% of the CA variants are found to cause a decrease the mass, while slightly less for the CI variants approximately 50% of the variants are shown to have a decrease in mass. The CA variants are found to have more variants that decrease the dipole moment than those observed in the CI variants, 75% and 40% respectively. Fewer of the CA variants are seen to cause a decrease in the pI (40%) but there are only slightly more CI variants that cause a decrease in the pI (55%). Mostly, we can see the most distinguishing feature with greatest change or difference or influence upon substitution is the dipole moment. A decrease in the dipole moment can allow for more lipophilic nature of the protein. It is more common to see lipophilic residues to be buried in the core of a structure with very little to no interaction with the solvent, though the presence of lipophilic surface sites exists.

We can also see the majority of substitutions resulting in a CI variant occur in the core and binding regions and are found to belong to a β -sheet or coil SSE. However, there are a few CI variants located with substitutions occurring in the hinge and elbow region as well as being part of an α -helix or turn element. The majority of substitutions resulting in a CA variant also occur in the core region of the structure located within a β -sheet SSE. Additionally, there are a few CA variants that appear on the elbow within a coil SSE. There

is only one substitution leading to a CA variant that takes place in the binding region within a β -sheet SSE. HIV-1 protease is a β -sheet rich structure and it's believed that a β -sheet rich core can lead to better packing (Ansari 2021, Li 2013, Slack 2015). As residues pack closer together or are pushed apart due to realignment the amount of mobility they have can also increase or decrease. Likewise, the residues can have new or lose contacts with other residues that can result in changes in the allosteric network common to the WT.

Results and Discussion

Of the 99 amino acids observed in HIV-1 protease 107 variants were evaluated by Martinez ¹ et.al., of which 12 of them are found to cause an active variant. On the other hand, 56 of the 107 variants evaluated, with only 42 of 56 being unique residue numbers of the 99 amino acids are found to cause an inactive variant Martinez ¹. Of the two cases (active and inactive) there are 2 shared residue numbers, residues K20 and I64, that result in dysfunctional variants. Overall, in this study the number of substitutions leading to an inactive variant is greater than that of having an active variant resulting an imbalanced data set. This can be highly important for the regulation of HIV-1 protease^{78, 109, 116, 224}, though the activity of the variants towards other substrates is unknown.

Here, RMSF is used to investigate the mobility changes of residues over the length of MD production run. Residues that have more mobility should be less constrained due to the packing of the structure and will result in a higher RMSF and vice versa. The figure

below shows the RMSF of each residue (only for C α) in chain A vs chain B (Figure 2.2) for the average of the 3 MD runs for the total MD production length of 72 ns for the WT structure. We see that most of the points fall near the linear regression line, which is expected as we deal with a homomeric protein dimer which has a C2 symmetrical behavior from the two identical chains. However, as the mobility increases, we can find a few residues with larger variations that deviate more from the linear fit line. These residues (49, 50, 51) are associated with the flap region of HIV-1 protease which is an important region of this protein structure that directly interacts and are essential for catalysis of the substrate. Note that one of these three residues, 50, is found to be a CI variant. To further analyze this data to understand the protein's activity upon a single point substitution, we compare the RMSF of each residue in WT structure to the one in variant structure.

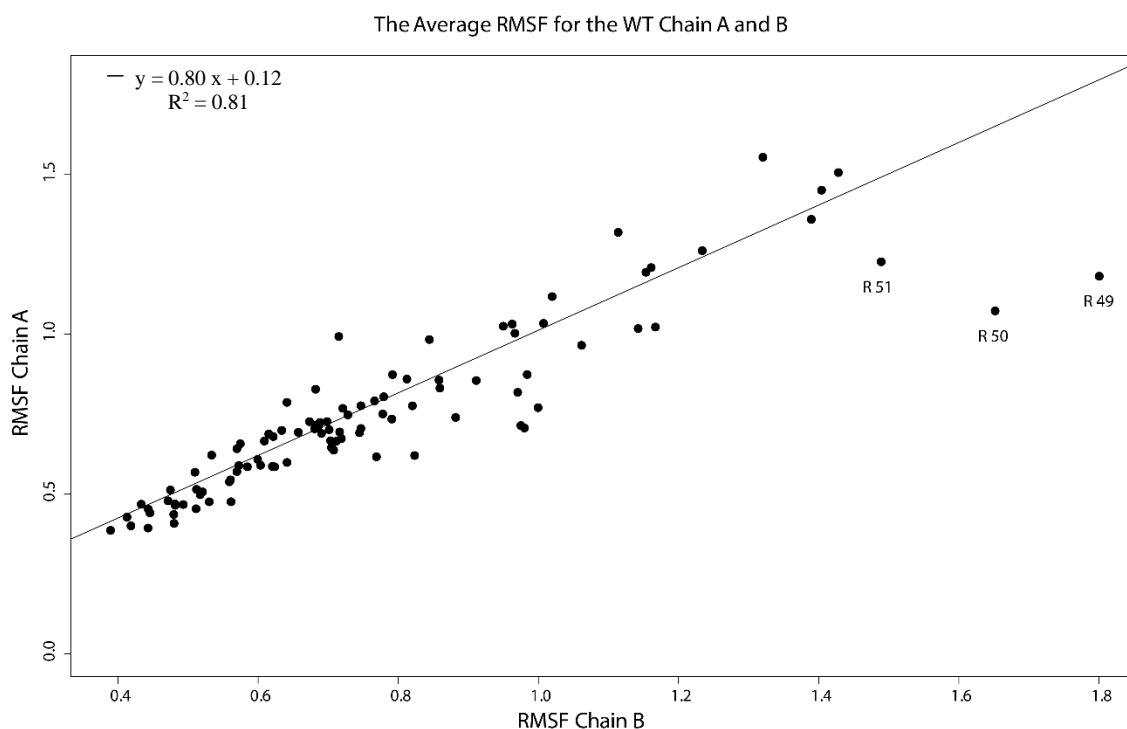


Figure 2.2: The figure above shows the average RMSF for chain A (y axis) and chain B (x axis) for the total 72 ns of MD production for each trajectory. The mean of all 3 trajectories is represented as the points with the best fit line shown by the equation in the top left.

The next set of figures, Figure 2.3 A&B, examines which residues of the variants are found to influence the mobility compared to the WT. Looking at the Figure 2.3 A, we can see the CA variants RMSF doesn't deviate much from the WT. Looking at residue numbers 49:51 that are associated with the flap region, we can see more often the CA variants lower the mobility of those residues. Observing the changes in mobility for residues 49:51 for the CI variants shows that though we see some of the CI variants lower the mobility, but more CI variants increase the mobility of those residues.

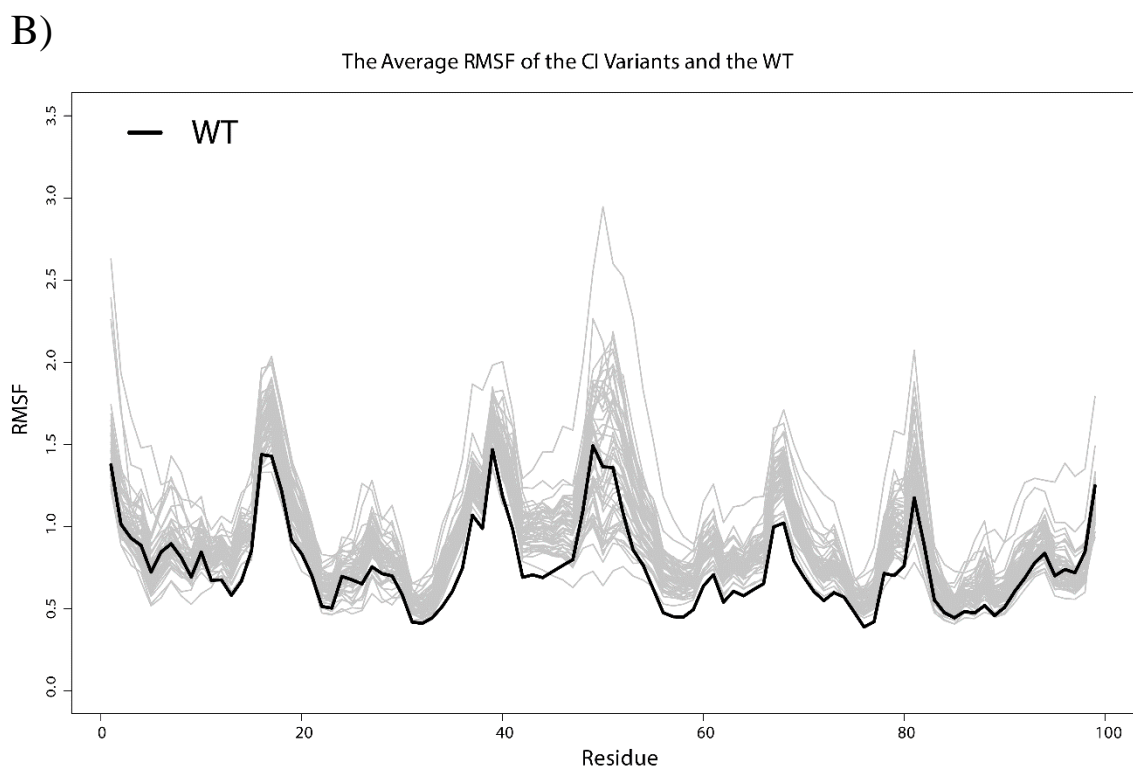
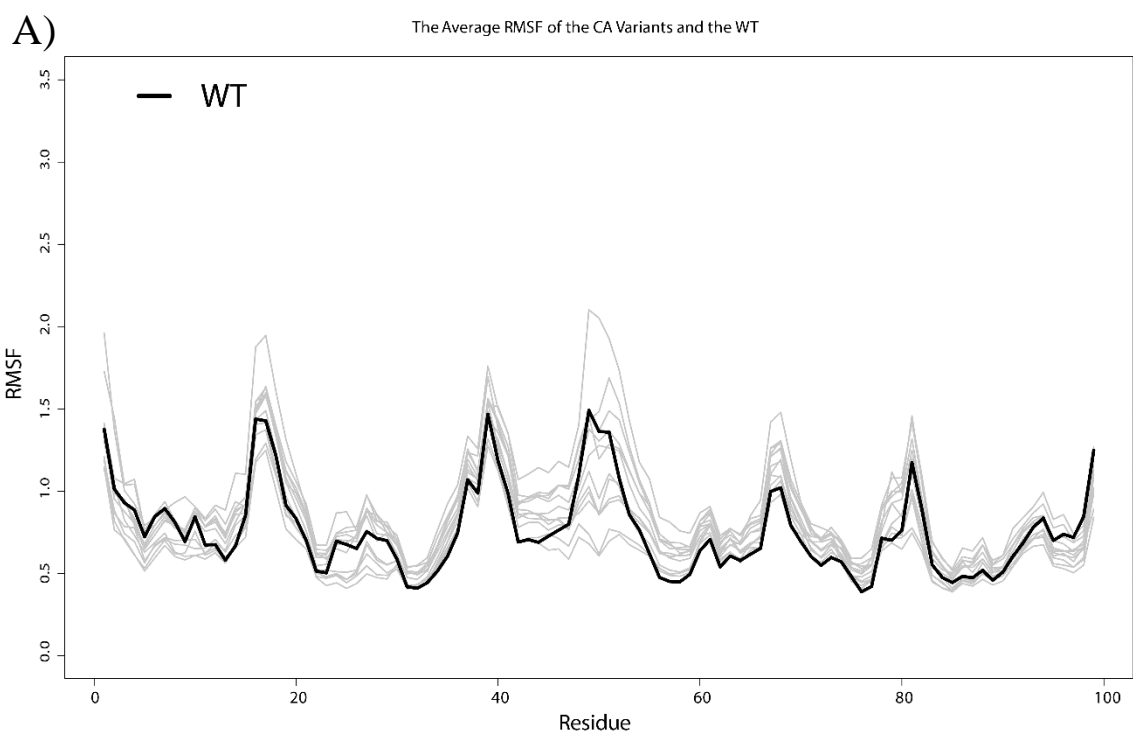
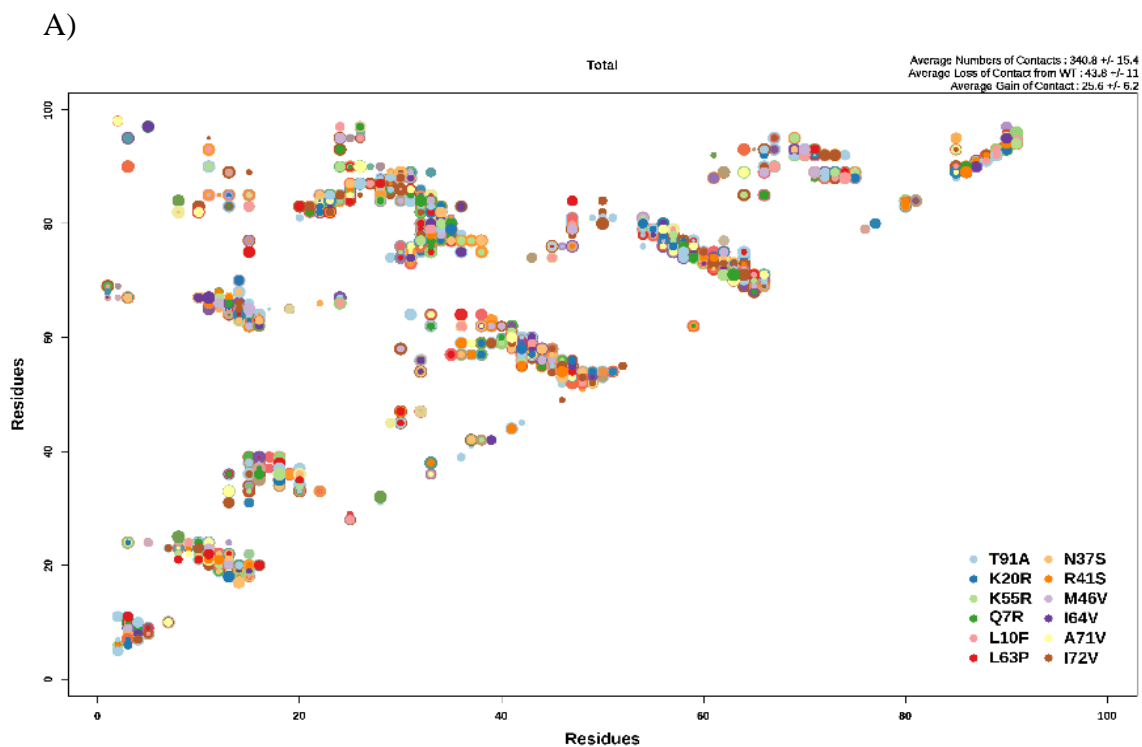


Figure 2.3 A&B: The figures above show the RMSF (y axis) for the CA variants (A) and CI variants (B) of the residue numbers (x axis) for the total length of production. The WT is represented as the black line and each variant is shown as gray (top left (A) & top right (B)).



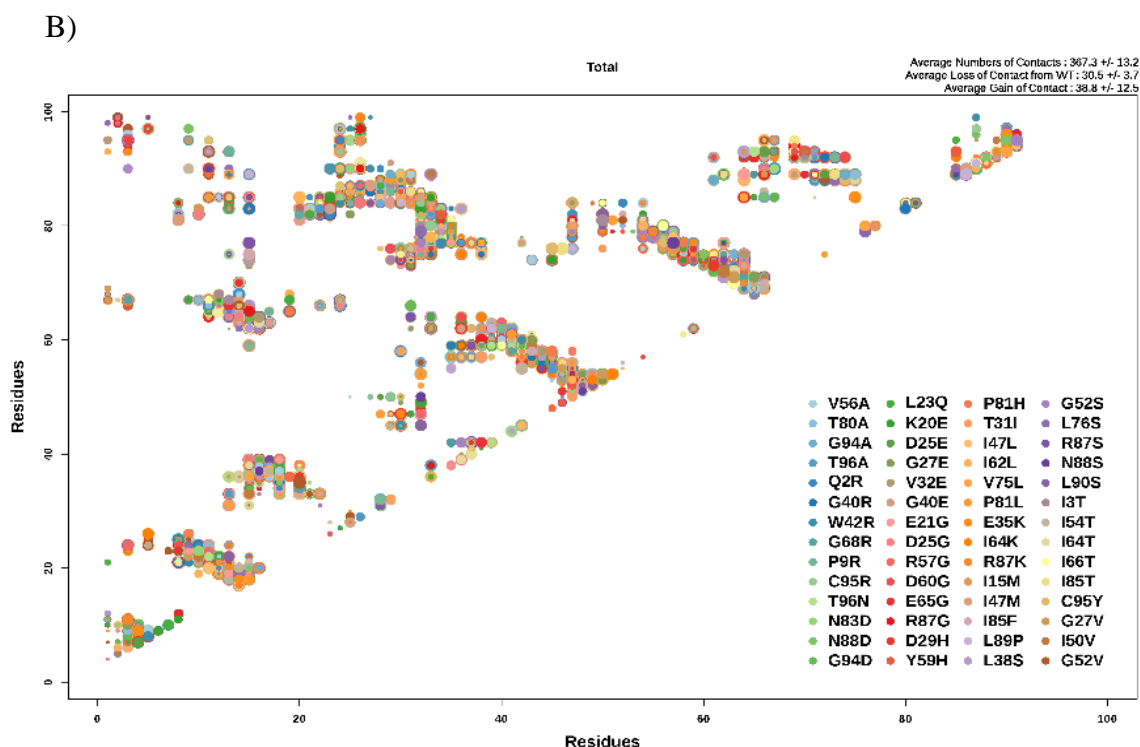


Figure 2.4 A&B: The figures above show the contacts that are observed in chain A for the CA variants (A) and CI variants (B). Each of the 99 residue numbers are shown on the x and y axis. The points represent contacts between 2 residues with the size of the point corresponding to the frequency observed throughout the production for the specific variant (see color legend).

Figures 2.4 A&B examines the total number of contacts that occur between non-neighboring residues ($r_i \pm 1$ or $r_i \pm 2$) throughout the MD production length for each variant structure and figure 2.5 A&B looks at all contacts that are seen to occur in 10% of the MD production length or more. However, through examination of higher frequency of contacts we observe differences in the contact maps of CA and CI variants, as well as differences in these maps compared to that of the WT, Figure S-2.3. There are a total of 359 non-neighboring contacts (542 with neighboring) observed in the WT (Figure S-2.3); however, it is observed that the total number of contacts in the CA variants decreases, an average of 341 non-neighboring contacts are seen compared to the 359 contacts found in the WT,

Figure 2.4A. Inversely, we see that the number of contacts found for most CI variants increases, an average of 367 contacts are found from the CI cases, Figure 2.4B. Though most of the residue contacts are preserved from what is seen in the WT, it is clear that the total number of contacts for a CA variant are less and the majority of the CI variants have more contacts. The observation of Figure 2.4 as well as RMSF results shows that in the case of CI variants, the residues tend to show higher mobility and higher number of contacts with regard to what is observed with the WT structure. As a result, it is likely the packing of the residues is unfavorable and causing the structures of the CI variants to be more rigid when bound with the substrate.

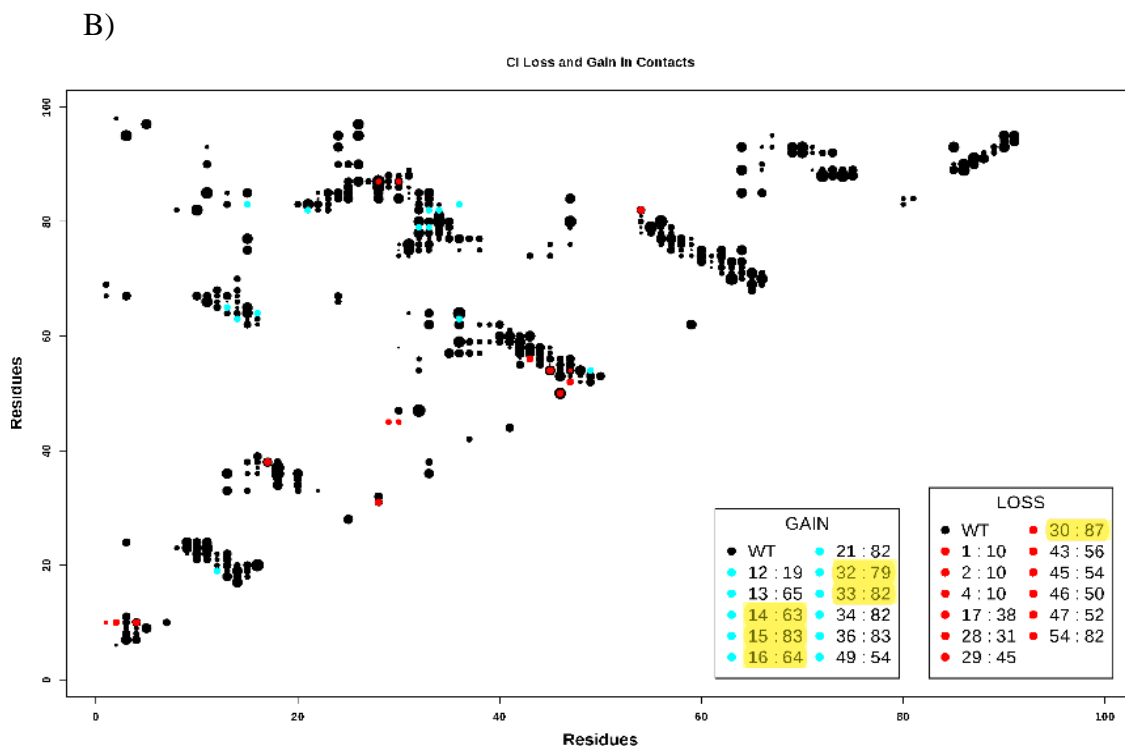


Figure 2.5 A&B: The figures above show the most common residue contacts that are gained or lost in chain A for at least 70% of the CA variants (A) and CI variants (B). Each of the 99 residue numbers are shown on the x and y axis. The points represent contacts between 2 residues with the size of the black points corresponding to the frequency observed throughout the production for the WT. Residues that are observed in the WT but not in the variant are considered a loss (red), inversely residues observed in the variant and not the WT are considered gain (blue). The size of the points is also scaled according to the average frequency observed from the corresponding variant case.

The next set of figures takes a deeper look at the impact of the gain or loss of contacts (Figures 2.5 A&B). This shows that the contacts of non-neighbor residues for the WT with the most common residues gained (blue) and lost (red) for the CA variants (A) and CI variants (B). As it was seen with Figures 2.4 A&B the CI variants have more contacts overall and we can see in Figures 2.5 A&B that the CI variants gain more but lose less contacts. The residue contacts that are exclusive for the CA or CI variant cases are

highlighted in the legend, Figures 2.5 A&B. It is observed that all of the common contacts gained for the CA variants are also represented in the CI variants, however there are several contacts that are gained in the CI variants not found in the CA variant cases. Inversely, we can see that there are several common loss of contacts represented in the CA cases not found in the CI variant cases, with there being only one novel loss of contact observed in the CI variants.

Figure 2.6 represents the residues of the CI variants that are found to have common gain of contact. We see that residue positions 14 and 16 make contacts with 63 and 64 found within adjacent beta sheets, whereas the contact between residues 15 and 83 is buried within the core of the structure, Figure 2.6. The distance from any of the sidechains for the residues ranges between 6.3–7.9 Å. Notably the cutoff used to identify whether residues contact occurs was 3 Å, Figure S-2.3. We can see that the contacts found elsewhere are mostly located in coils within the core of the structure, Figure 2.6. In particular, residues 32-33 make contact with residues 79 and 82, respectively, and from the measurements observed in the WT we see the distance between them ranges from 5.6–9.5 Å. It can be observed that in the WT and given the orientation of these residues it takes a significant rearrangement for most/all of these contacts to occur.

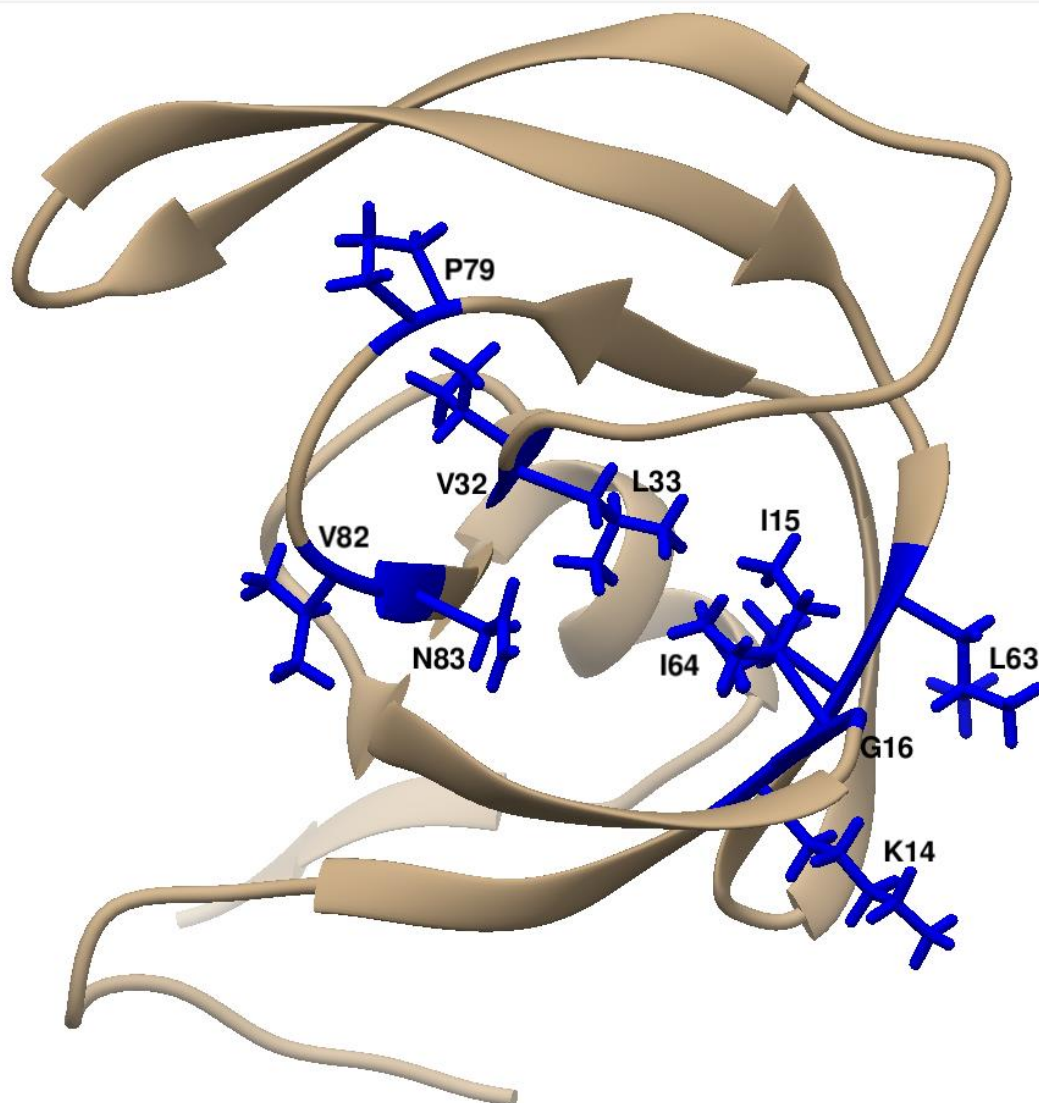


Figure 2.6: These figures identify the structural orientation of the residues found to have novel gain of contact for the CI variants.

Considering the novel loss of contacts observed in the CA variants we can see that there are 3 main regions that are impacted (the loop, core, and cantilever). It can be assumed that since these loss in contacts results in a CA variant, exclusively, these residue contacts might be least supportive towards an allosteric mechanism. While we can see that the only novel loss of contact observed in the CI variant cases are between residue numbers 30:87,

which falls within the binding site region. It is likely that given the importance of these residues and their influence in coordinating the substrate for catalysis they might be an intricate component towards influencing the activity of a variant.

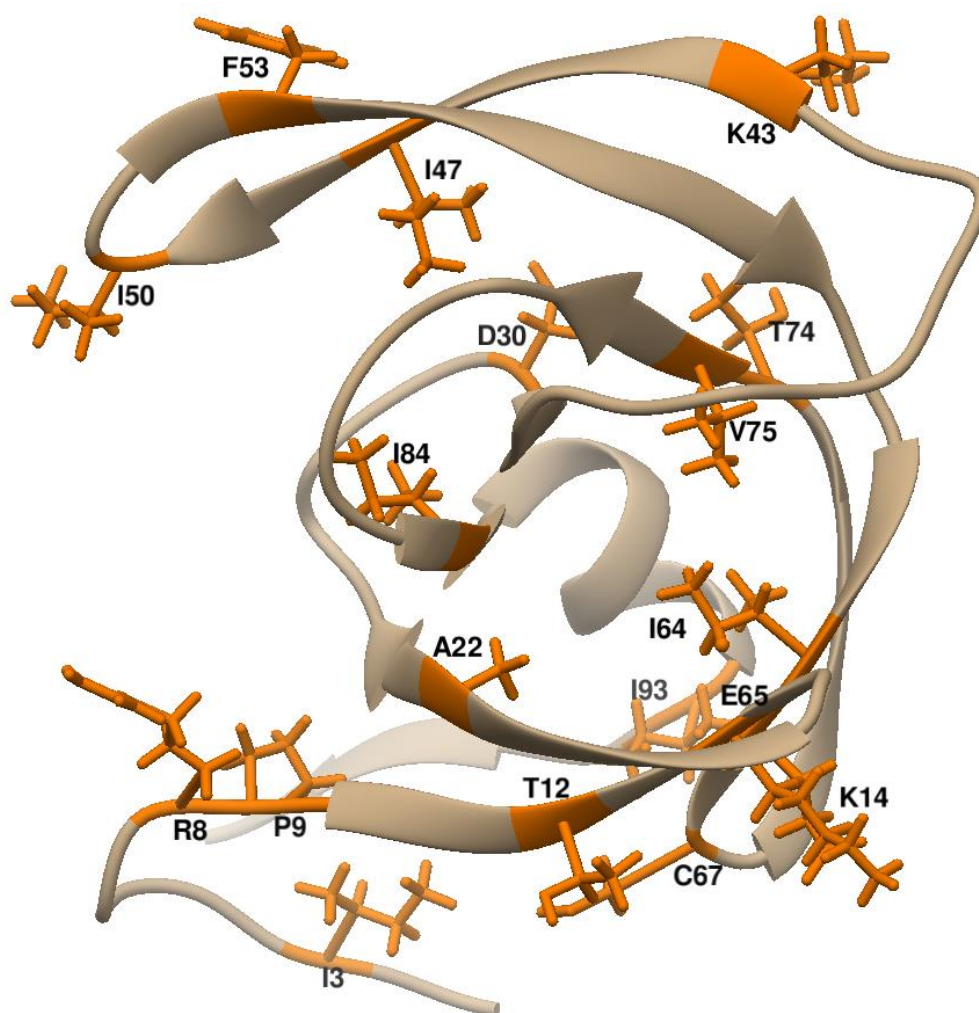


Figure 2.7: These figures identify the structural orientation of the residues found to have novel loss of contact for the CA variants.

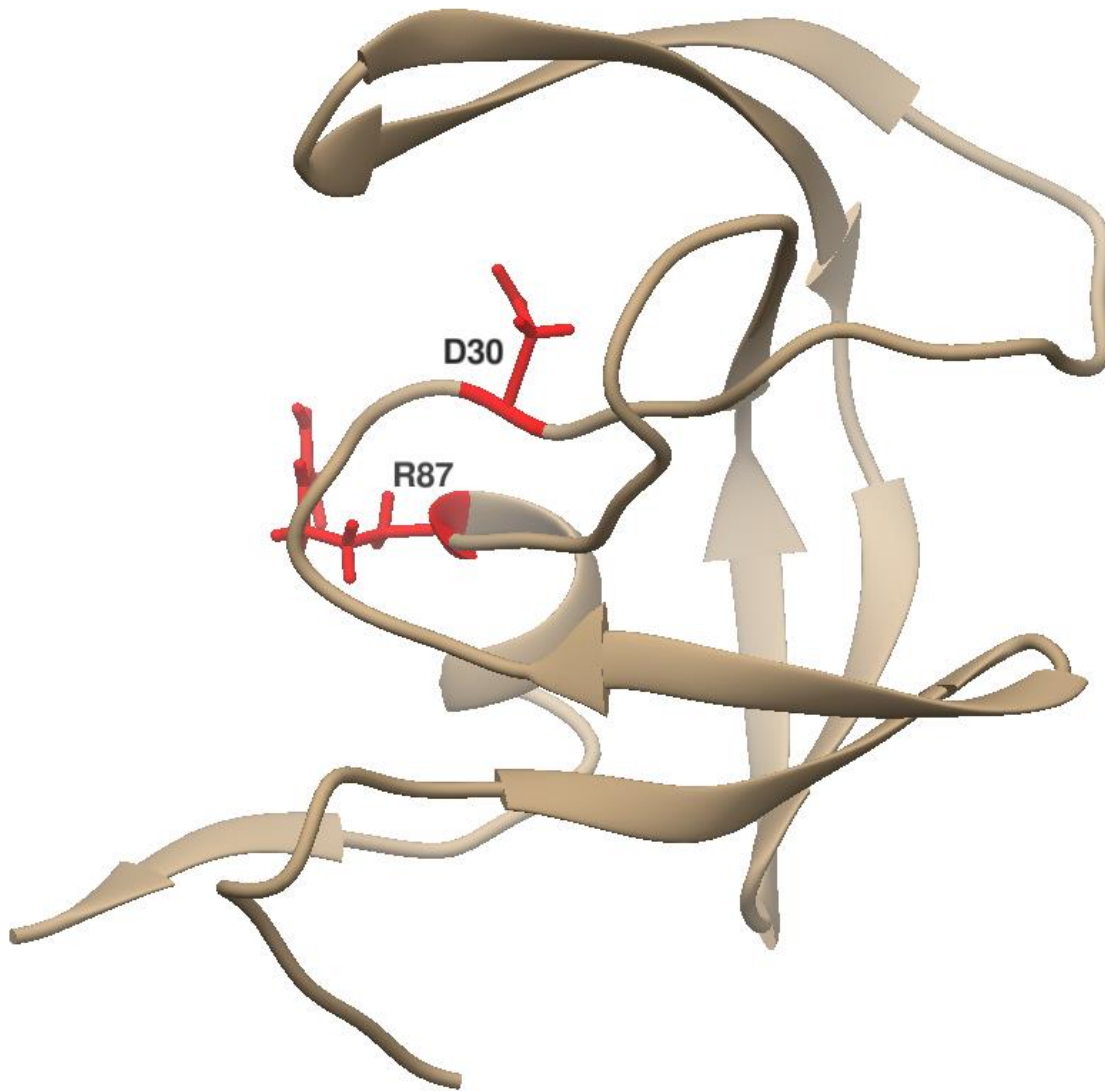


Figure 2.8: These figures identify the structural orientation of the residues found to have novel loss of contact for the CI variants.

We've examined the gain or loss of contact as well as the mobility in great detail, but we also want to understand if the differences observed influence the folding and/or binding free energy change upon single point substitution for the WT. The folding and binding free energy of each variant was scrutinized extensively using MMPBSA

calculations (eq. 2.7) as well as these software webserver; Prodigy, DUET, INPS, Eris, and DDGun^{105, 110, 227-235}, over the entire length of the MD production as well as calculations using the last frame of the MD production on remote web servers. The first component we will look at is the folding free energy change upon substitution (ddG). The total production length of a single monomer (chain A) of the WT and each variant was assessed by computing MMPBSA in order to approximate the change in free energy (dG) for each structure as well as the folding free energy change upon substitution (ddG). MMPBSA (eq. 2.7) is composed of three terms, with the first taking into account the translational mechanical motion with respect to classical mechanics, $F = ma$ ^{115, 236-238}. The second term looks at the energy impact from the electrostatic free energy associated with the polar solvent using Poisson-Boltzmann equation^{115, 236-238}. The last term considers the influences given by non-polar solvent with regards to the solvent accessible surface area (SA)^{115, 236-238}.

Two approaches were taken to understand the appropriate methodology to evaluate the change in free energy from the entire production using MMPBSA. The most common methodology used to calculate the folding free energy change upon substitution (ddG) is done by getting the average change in free energy (dG) for each structure or only using the last frame of production. Here we institute a third novel more dynamic measure of the folding free energy change upon substitution (ddG) from the output of the entire production. We consider the difference between the maximum and minimum change in free energy from the entire length of production, as a way to measure the range of change

of the Michaelis complex of the protein with the bound substrate. This method was approached since the change in energy of the system is dynamic, such that the change in MM/PB/SA can vary significantly throughout the production length. The concept of using the average change in free energy (dG) for the length of production provides reasonable coverage of the dynamic nature of a MD run as well.

Looking at Table 2.2, we see that the average change in free energy is $dG(1)$ and the average change in folding free energy upon substitution is $ddG(1)$, whereas, the change in free energy, $dG(2)$, and the change in folding free energy upon substitution, $ddG(2)$, are the difference (maximum and minimum) for the entire MD production length. In general, since we are only considering the change in folding free energy caused by a single point substitution the overall change is not expected to be extreme, Table 2.2 particularly with regard to $ddG(2)$. Distinctly, most of the variants in either case (CA or CI) are found to be stabilizing with regards to $ddG(2)$, Table 2.2. Conversely, ddG^1 shows only 50% of the CA and ~30% of the CI are stabilizing and likewise when we take the average of both methods ($ddG(1)$ and $ddG(2)$) together shows about the same result, Table 2.2. However, the overall nature of the results of $ddG(2)$ alone corresponds better with the observations of the change in free energy upon substitution from the webserver^{105, 110, 227-235}, Table 2.2 and Table 2.3.

Table 2.2: Change in folding free energy (dG) and change in folding free energy upon substitution (ddG) using MMPBSA for method 1 (average) and method 2 (difference).

Folding Free Energy upon Substitution ddG							
Variant	dG ¹	dG ²	<dG ^{1,2} >	ddG ¹	ddG ²	<ddG ^{1,2} >	Activity
WT	119.37	24.88	72.13	0	0	0	1
Q7R	137.57	22.52	80.04	18.2	-2.37	7.91	1
A71V	119.23	23.11	71.17	-0.14	-1.77	-0.96	1
L10F	138.42	23.19	80.8	19.05	-1.7	8.67	1
N37S	122.69	23.34	73.01	3.31	-1.55	0.88	1
M46V	143.97	23.34	83.66	24.6	-1.55	11.53	1
K55R	115.45	23.45	69.45	-3.92	-1.43	-2.68	1
L63P	109.85	23.46	66.66	-9.52	-1.43	-5.47	1
I64V	106.21	24.35	65.28	-13.16	-0.53	-6.85	1
R41S	132.14	25.02	78.58	12.77	0.13	6.45	1
I72V	103.85	25.21	64.53	-15.52	0.32	-7.6	1
K20R	123.26	25.55	74.41	3.89	0.67	2.28	1
T91A	109.64	25.83	67.73	-9.73	0.94	-4.39	1
G52V	140.84	21.13	80.98	21.46	-3.76	8.85	0
I54T	149.49	22.27	85.88	30.12	-2.62	13.75	0
I47M	146	22.41	84.2	26.62	-2.47	12.07	0
G27E	139.3	22.42	80.86	19.93	-2.47	8.73	0
L89P	130	22.46	76.23	10.62	-2.43	4.1	0
L76S	135.41	22.56	78.99	16.04	-2.32	6.86	0
P81H	130.61	22.58	76.59	11.23	-2.31	4.46	0
N83D	120.52	22.84	71.68	1.15	-2.04	-0.45	0
L38S	120.7	22.84	71.77	1.33	-2.04	-0.36	0
Q2R	120.24	22.88	71.56	0.87	-2	-0.57	0
P9R	126.67	22.96	74.82	7.3	-1.92	2.69	0
D25G	124.57	22.97	73.77	5.2	-1.91	1.64	0
G68R	135.95	23.18	79.57	16.58	-1.7	7.44	0
D29H	127.22	23.3	75.26	7.85	-1.59	3.13	0
R87S	131.33	23.3	77.31	11.96	-1.59	5.18	0
P81L	136.32	23.38	79.85	16.95	-1.51	7.72	0
I64T	123.43	23.39	73.41	4.06	-1.5	1.28	0
T80A	150.05	23.41	86.73	30.67	-1.47	14.6	0
N88S	127.03	23.52	75.27	7.66	-1.37	3.15	0
R87K	122.28	23.55	72.92	2.91	-1.33	0.79	0
R87G	151.92	23.55	87.74	32.54	-1.33	15.61	0
V75L	142.54	23.63	83.08	23.17	-1.26	10.95	0
G40E	155.76	23.65	89.71	36.39	-1.24	17.58	0
E65G	117.82	23.7	70.76	-1.56	-1.19	-1.37	0
C95Y	128.02	23.72	75.87	8.64	-1.17	3.74	0
T96N	115.94	23.77	69.86	-3.43	-1.12	-2.27	0
L23Q	118.6	23.89	71.25	-0.77	-0.99	-0.88	0
K20E	112.18	23.94	68.06	-7.19	-0.94	-4.07	0
T31I	133.7	23.97	78.84	14.33	-0.92	6.71	0
C95R	129.1	24	76.55	9.72	-0.88	4.42	0
V56A	126.7	24.01	75.35	7.32	-0.88	3.22	0
I15M	124.97	24.03	74.5	5.6	-0.86	2.37	0
T96A	117.99	24.08	71.03	-1.39	-0.81	-1.1	0
I85T	116.11	24.08	70.1	-3.26	-0.8	-2.03	0
E35K	121.11	24.27	72.69	1.74	-0.61	0.56	0
I47L	115.78	24.29	70.04	-3.59	-0.59	-2.09	0
D25E	121.78	24.44	73.11	2.41	-0.45	0.98	0
E21G	121.88	24.6	73.24	2.51	-0.29	1.11	0
G94D	109.86	24.65	67.25	-9.52	-0.24	-4.88	0
I62L	113.06	24.68	68.87	-6.31	-0.2	-3.26	0
L90S	122.66	24.69	73.67	3.28	-0.2	1.54	0
G40R	112.47	24.85	68.66	-6.91	-0.04	-3.47	0
I3T	115.78	24.87	70.33	-3.59	-0.01	-1.8	0
I66T	115.26	24.87	70.07	-4.11	-0.01	-2.06	0
Y59H	118.41	24.88	71.65	-0.96	0	-0.48	0
G52S	105.31	24.89	65.1	-14.06	0	-7.03	0
G27V	121.22	24.93	73.08	1.85	0.05	0.95	0
I50V	121.39	24.96	73.18	2.02	0.07	1.05	0
R57G	129.27	25.36	77.32	9.89	0.48	5.19	0
I64K	106.24	25.51	65.88	-13.13	0.63	-6.25	0
W42R	103.03	25.54	64.28	-16.35	0.65	-7.85	0
N88D	123.42	25.88	74.65	4.04	0.99	2.52	0
D60G	118.85	26.05	72.45	-0.52	1.17	0.32	0
G94A	134.12	27.03	80.57	14.75	2.14	8.44	0
I85F	113.54	27.54	70.54	-5.83	2.66	-1.59	0
V32E	120.42	28.23	74.32	1.04	3.34	2.19	0

Model	Active	Inactive
ddG ¹	0.5	0.32
ddG ²	0.83	0.82
<ddG>	0.5	0.36

"Fraction of variants found to be stabilizing"

The next table reports the change in folding free energy upon substitution from several webserver^{105, 110, 227-235} using the last snapshot from the production, Table 2.3. The data collected from the DUET, INPS, Eris, and DDGun webserver^{105, 110, 227-235} affirms the magnitude and scale identified from ΔG^2 calculated from MMPBSA, Table 2.2 and Table 2.3. However, we do see that there is a distinct trend where all of the webserver predict that more often the CA variants are less stabilizing than the CI variants, Table 2.3 B. The change in free energy upon substitution (ΔG) calculations shown in Table 2.2 and 2.3 are informative, in that, though a monomer (chain A) of a variant may be CI the fold stability of the variant is greater than the WT. It also constituents that though the CI variants may be inoperable towards the given target they may be functional towards others and ultimately lower the drugability of functionally active variants. Further, it supports that many of these variants may be locked in a specific conformation (open or closed) that is stable yet unable to bind the substrate or become locked once a substrate is bound.

Table 2.3: Change in folding free energy upon substitution (ddG) using the DUET,
INPS, Eris, and DDGun webserver^{105, 110, 227-235}

Variant	mCSM	SDM	DUET	DDGun	"-Eris"	INPSMD3d	Average	Activity
L63P	-0.57	-2.12	-0.71	-0.3	-6.69	-0.39	-1.8	1
I64V	-1.73	-2.38	-2.06	-1	0.1	-0.34	-1.24	1
R41S	-0.3	-0.67	-0.27	-0.1	-4.6	-0.36	-1.05	1
M46V	-0.96	0.86	-0.23	-0.1	-4.04	0.02	-0.74	1
I72V	-0.56	0.31	-0.07	-0.5	-2.5	0	-0.55	1
L10F	-0.75	-0.15	-0.69	0	-1.71	0.62	-0.45	1
K20R	-0.73	-0.21	-0.39	0.7	-0.31	-0.27	-0.2	1
Q7R	0.22	0.63	0.58	-0.1	-2.31	0.3	-0.11	1
N37S	-0.22	-0.43	0.01	-0.1	0.97	-0.01	0.04	1
K55R	-0.39	-0.14	-0.02	0.4	0.52	-0.09	0.05	1
T91A	-0.49	1.98	0.33	-0.2	0.35	-0.16	0.3	1
A71V	-0.7	0.95	-0.07	1.4	1.17	0.84	0.6	1
I64K	-2.37	-3.53	-2.66	-5.4	-7.13	-3.15	-4.04	0
L76S	-3.17	-3.89	-3.5	-5.1	-3.21	-2.77	-3.61	0
I66T	-2.58	-3.36	-2.88	-3.6	-5.11	-2.87	-3.4	0
I85T	-2.78	-3.36	-3.08	-3.9	-3.79	-3.26	-3.36	0
L89P	-1.56	-4.41	-2.25	-4.1	-2.83	-3.25	-3.07	0
I64T	-2.82	-3.36	-3.11	-4	-1.57	-3.02	-2.98	0
V56A	-2.36	-2.66	-2.81	-3.2	-4.69	-1.92	-2.94	0
V32E	-2.65	-1.99	-2.78	-4.2	-2.13	-2.7	-2.74	0
L90S	-1.96	-3.57	-2.3	-3.3	-1.77	-2.91	-2.63	0
I85F	-1.76	-1.54	-2.01	-1.4	-6.23	-1.08	-2.34	0
R57G	-2.41	0.19	-2.09	-2.3	-4.67	-1.96	-2.21	0
C95Y	-0.48	-0.18	-0.45	-1.1	-10	-1	-2.2	0
L38S	-1.52	-2.8	-1.75	-4.1	0.35	-2.92	-2.12	0
L23Q	-1.62	-1.42	-1.52	-3	-2.84	-1.42	-1.97	0
E65G	-1.22	0.56	-0.85	-1.4	-7.56	-1.09	-1.93	0
I54T	-1.04	-0.79	-0.76	-1.5	-5.3	-1.9	-1.88	0
C95R	0.46	0.75	0.7	-1.7	-10	-1.04	-1.8	0
R87G	-1.19	-1.39	-1.39	-1.8	-1.58	-1.18	-1.42	0
R87S	-1.26	-2.05	-1.56	-1.2	-1.39	-0.98	-1.41	0
I3T	-0.5	-1.46	-0.43	-0.9	-3.88	-0.94	-1.35	0
G52S	-0.58	-3.19	-0.85	-0.5	-1.97	-0.67	-1.29	0
N83D	-1.31	0.11	-1.01	-1	-3.37	-1.09	-1.28	0
E35K	-0.01	-0.59	0.2	-0.1	-6.61	-0.29	-1.23	0

Model	Active	Inactive
mCSM	0.92	0.96
SDM	0.58	0.7
DUET	0.75	0.84
DDGun	0.75	0.98
"-Eris"	0.58	0.7
INPSMD3d	0.67	0.91
Average	0.67	0.93

"Fraction of variants found to be stabilizing"

I15M	-0.85	-1.91	-1.14	-2	0.48	-1.74	-1.19	0
I62L	-1.02	-0.64	-0.82	-1.1	-3.22	-0.28	-1.18	0
R87K	-1.12	-0.57	-0.93	-0.8	-2.66	-0.87	-1.16	0
K20E	-0.14	-0.23	0.18	-0.7	-5.09	-0.89	-1.15	0
I47L	-0.82	-0.56	-0.57	-1.1	-2.58	-0.2	-0.97	0
G94D	-0.64	-2.92	-0.96	-0.7	0.03	-0.42	-0.93	0
W42R	-0.84	-0.36	-0.64	-2.4	0.27	-1.6	-0.93	0
N88D	-1.55	-0.68	-1.4	-0.5	-0.84	-0.28	-0.88	0
Y59H	-2.11	-1.07	-2.08	-2.7	4.27	-1.17	-0.81	0
V75L	-0.47	-1.5	-0.44	-1.5	-0.07	-0.76	-0.79	0
T80A	-0.72	0.22	-0.47	-1.4	-1.36	-0.78	-0.75	0
G27E	-0.94	0.17	-0.52	-0.8	-1.5	-0.9	-0.75	0
G52V	-0.29	-1.52	-0.34	0	-0.83	-1.44	-0.74	0
I47M	-0.64	-0.97	-0.66	-1.5	0.09	-0.61	-0.72	0
N88S	-0.74	-0.9	-0.6	-0.7	-0.67	-0.56	-0.7	0
G94A	-0.15	-2.71	-0.53	-0.3	1.13	-0.83	-0.57	0
G40R	-0.35	-1.5	-0.3	-0.3	-1.41	0.75	-0.52	0
D25E	-0.75	-0.24	-0.53	-0.4	-1.68	0.69	-0.48	0
D29H	-0.38	0.04	-0.23	-0.4	-1.23	-0.54	-0.46	0
E21G	-0.68	0.28	-0.33	-0.5	-0.53	-0.78	-0.42	0
T31I	-0.23	0.95	0.22	0.2	-2.6	-0.53	-0.33	0
G27V	-0.38	0.15	-0.01	-0.8	0.89	-1.64	-0.3	0
G40E	-0.89	-1.31	-0.79	-0.5	1.55	0.36	-0.26	0
P81L	-0.41	0.84	0.11	-0.2	-0.8	-0.71	-0.2	0
I50V	-0.65	-0.09	-0.35	-0.1	0.95	-0.58	-0.14	0
T96N	-0.18	-0.06	0.11	-0.4	0.09	-0.19	-0.11	0
D60G	-0.69	1.14	-0.12	-1	0.8	-0.68	-0.09	0
T96A	-0.18	0.15	0.08	-0.4	0.35	-0.41	-0.07	0
G68R	-0.39	-2.39	-0.46	-0.3	3.32	-0.18	-0.07	0
P9R	-0.51	1.12	-0.12	-1.6	2.15	-0.86	0.03	0
D25G	-0.81	0.34	-0.57	-1.2	2.16	0.71	0.11	0
Q2R	0.11	0.87	0.5	0	-0.97	0.35	0.14	0
P81H	-0.03	0.91	0.31	-0.3	2.34	-0.22	0.5	0

We also used Prodigy^{105, 228, 229} as well as MMPBSA to calculate the change in binding free energy upon substitution for the enzyme substrate complex of the WT and each variant, Table S-2.1. A notable comparison of the change in free energy (ΔG) values calculated using prodigy with the difference method used with MMPBSA, we do see a slight over estimation of the change in free energy from the difference method. Additionally, prodigy predicted that all of the CA variants and most of the CI variants are destabilizing. Whereas the difference method used to calculate the change in free energy upon substitution ($\Delta\Delta G^2$) shows most the CA variants are destabilizing while almost half of the CI variants are stabilizing.

Conclusion

It is not directly apparent how the chemical and physical properties influence the changes in a protein without characterizing the dynamic nature of a structure. We observed that many of the changes that occur in the CI variants of HIV-1 protease impact the mobility compared to the CA variants and WT. The strongest evidence is seen with regard to the gain and loss of contacts. We can see that the fewer number of contacts provides the CA variants greater or comparable activity to the WT, while the CI variants gain more and lose less contacts resulting in a greater number of contacts compared to the WT and CA variants. It is likely that the overall increase in contacts for the CI variants affects the conformational states that these residues transition between, whether there is more states or restriction of states.

It is also observed that the change in free energy upon a single point substitution for most of the variants are stabilizing, either CA or CI variants. Though there are mixed results with regards to the folding/binding free energy change upon substitution, it is more adequately characterizing the flexibility and rigidity of the structure. It is suggestive that the CI variants, based on the influence of the contacts observed, are stable but locked in various inoperable conformations. The most interesting consensus is shown from the webserver^{105, 110, 227-235}, such that more often the CI variants are found to be more stabilizing. Looking at this from a scale of flexible to rigid, “stable” would be more rigid and the overall increase in contacts confirms this assessment.

CHAPTER III

THE IMPORTANCE OF ANTICORRELATION ON THE FUNCTION OF HIV-1 PROTEASE VARIANTS AND PREDICTING THE FUNCTION OF THE PROTEIN UPON SINGLE POINT SUBSTITUTION USING A ROBUST MACHINE LEARNING ALGORITHM AND DYNAMIC CROSS CORRELATION

Introduction

The evolutionary pathways and outcomes of an enzyme are significantly regulated by constraints imposed from the deleterious effects of amino acid substitutions on its functional properties^{108, 239, 240}. And therefore, identification and characterization of these constraints are critical for the efficiency and function of an enzyme as well as the categorization of variants that can be potentially drug resistant targets^{108, 239, 240}. Past and recent experimental and theoretical studies suggest that the characteristic dynamics of an enzyme are distinctive attributes that are evolutionarily selected features used to control their catalytic activity^{66, 126, 128, 241-248}. Essentially, for an enzyme to retain its catalytic activity, the mobility of the enzyme should remain the same or closest to what is observed in the WT. It has been further suggested that substitutions that have deleterious effects on the catalytic activity can alter the dynamics of an enzyme, imposing the idea that dynamics

can act as the constraint on its evolution^{66, 126, 128, 241-248}. The role of dynamics as a constraint on the evolution of an enzyme has not been very well characterized.

The amino acid sequence that makes up a protein structure is resourcefully unique and so are the low frequency side chain fluctuations each residue encounters as well as the global perturbations from each domain of the structure²⁴⁹. The term dynamics can be used to broadly describe these collective low and high frequency motions that have been evolutionarily adapted and optimized to provide specific vibrational modes to improve the interaction and function of a protein⁶⁸. Several NMR (Nuclear Magnetic Resonance) and theoretical studies examine how a protein's dynamics can be used to identify and interpret molecular recognition, allosteric regulation, and enzyme catalysis^{74, 128, 250-253}.

One way to computationally characterize the effect of motion on the activity of an enzyme has been to analyze the covariance of displacement of each residue pair within the enzyme¹²⁸. Molecular dynamic (MD) simulations have been shown to capture most of the relevant correlated motions of residue interactions, including those deemed as the low frequency harmonic motion from the side-chain fluctuations that can support the interpretation of allosteric modulation^{122, 194-198}. Analyzing the correlated motions from a pair of residues using MD simulations has been frequent practice to study the role of a protein dynamics^{86, 165, 254-257}. These linearly correlated motions observed from the coordinates of MD simulations show the positively correlated motions are mainly a side-effect of the secondary structures but can also result from long range coupling within the

protein^{129, 137, 258, 259}. However, most of the negatively correlated (anticorrelated) motions are not regarded as the result of a trivial relationship between residues and have been hypothesized to be an important dynamic feature linked to the activity of an enzyme¹²⁸. It has been suggested that the features that are important for the function of an enzyme are evolutionarily selected based on their anticorrelated motions with regard to specific residues^{117, 260-262}. Substitutions of the residues that contribute to the disruption of the anticorrelated regions are viewed as interfering with the activity of an enzyme¹²⁸. Therefore, it is postulated that a change in linear correlation can be considered as a parameter to address the impact a substitution has towards the function of any enzyme¹²⁹.

The human immunodeficiency virus type 1 (HIV-1) protease is a virally encoded enzyme responsible for the cleavage of Gag and Gag-Pol poly proteins at eleven sites of the immature AIDS virus playing an important role in its maturation process for viral replication¹⁸². The HIV-1 protease has a total of 198 amino acid residues that makes up the robust C2 symmetrical homodimer with two identical 99 amino acid residue chains¹⁸³⁻¹⁸⁵. Symmetry has been observed to contribute to the evolutionary stability and cooperative interactions between subunits of many oligomeric proteins, like that of HIV-1 protease^{54, 186-190}. HIV-1 protease is one of the major targets towards AIDS treatment, due to its key role in the viral life cycle, as well as the scrutinized during drug trials of protease inhibitors¹⁹¹⁻¹⁹³. However, the inefficient replication process of HIV-1 virus causes different mutational variants of HIV-1 protease to be rapidly produced allowing for the competition of functional variants. Some of these mutational variants can cause resistance against the

drug binding of protease inhibitors making them preferentially selected for viral replication in the presence of a protease inhibitor drug ¹⁹¹⁻¹⁹³. This process of selecting different mutational variants of HIV-1 protease is looked on as enzymatic evolution ^{159, 263}.

In the current study, the role of how the dynamics can act as a constraint on the evolution of HIV-1 protease is sought by identifying how the highly correlated (both positively and negatively) residues from an MD simulation impact the known in-vitro experimental catalytically active and inactive variants (CA and CI, respectively), Figure 3.1. The publication of the in-vitro experimental catalytic activity for 107 HIV-1 protease variants as measured towards the cleavage of the matrix-capsid (MA/CA) polypeptide domain has prompted us to ask whether (in this larger dataset) correlated motions can be meaningfully associated with the protein activity¹. Experimental data published by Martinez et. al.¹ demonstrated the effect of a single point substitution on the catalytic activity of HIV-1 protease towards the cleavage of MA/CA. The information provided by Martinez et.al.¹ used in conjunction with the recently determined structure of HIV-1 protease in the bound state with the MA/CA peptide substrates are sourced to explore the correlated motions within the relevant Michaelis complex ²⁰⁸. Each variant is described by having a single amino acid substitution repeated in each chain of the C2 symmetrical homodimer, Figure 3.1. The theoretical background of how residue dynamics acts towards improving or disrupting the catalytic efficiency of enzymes is not well established. Deriving the covariance between each residue pair of the HIV-1 protease and its variants

determined from a MD simulation is a suitable method to evaluate the relationship of dynamics and its association to in-vitro experimental activity.

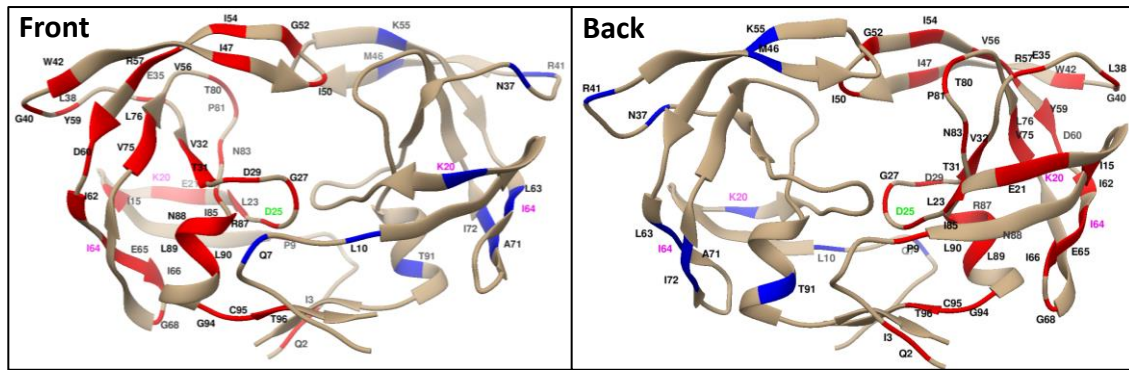


Figure 3.1: The figure shown above is a front and back view of the HIV-1 protease (PDB id: 1hwx). The colored mutant residue positions are highlighted in a single chain, though each substitution was applied to both chains, to show the location of the CI variant and CA variant residue positions (red and green respectively). Of the 99 common residues that make up each chain, there are 42 unique positions are found to cause an inactive variant and 12 positions are found to cause activity equal or greater than the WT structure¹.

Inspection of the in-vitro experimental data¹ presents 68 variants of the HIV-1 protease with 12 characterized as having activity equal or greater than the WT as well as 56 (42 unique residue positions) variants characterized as completely inactive compared to the WT structure, Figure 3.1. The amount of experimental data¹ accompanied with the use of advanced statistical modeling methods makes it possible to elicit if any statistically significant results can be observed from the impact of the correlated motion on an enzyme's activity. Modeling the HIV-1 protease WT and variant structures in the bound state with one of its natural substrates can offer insight of a bridge between the dynamics and the function of the robust symmetrical enzyme.

In the current study, we examine whether the residue positions found to possess significant correlations have any influence on the experimentally determined activity¹ (variant activities). In the following sections we look to uncover whether or not the correlated motion of residue pairs found to exhibit anti-correlation act as a dominant feature with regards to the function of a protein. We developed detailed statistical methods that can be used to explore the evidence linking the correlated motions from each pair of residue positions to the activity of the variants. Recognition of any substantial association between the dynamics of specific residues and allosteric interactions requires a detailed examination of the contributing factors that can regulate the dynamic cross-correlation (DCC) and the function of HIV-1 protease. We weigh the regulatory nature that the dynamics has on the activity of the HIV-1 protease variants by using machine learning (ML) algorithms to solve the binary-classification problem with regards to the function of the variant. Furthermore, the data from the DCC for each variant is assessed by the ML algorithms to predict if a single point substitution will lead to the disfunction in the activity of a protein compared to WT. The approaches taken here can help experimentalist as well as theorists in the study of enzyme activity as well as gain better control of allosteric interactions.

Methods

Structure modeling, molecular dynamics simulation and covariance analysis:

The models of HIV-1 protease with the MA/CA substrate native and variant complexes were prepared using the following methodology. The structure of the MA/CA substrate and a tetra coordinated water molecule were extracted from a crystal structure (PDB id: 1KJ4)²⁰⁸ and placed into the binding pocket of the native structure (PDB id: 1HXW)²⁰⁹ that is described as having the same amino acid sequence consistent with the wild type (WT) structure²¹⁰. The ligand and water were placed into the binding pocket of the WT structure by superposition of the backbone atoms of its active site residues (within 4.5 Å of ligand) with the corresponding atoms of 1KJ4. The single point substitutions were created using the appended WT structure generate each variant by using the software called MODELLER¹⁴. As described in the literature the bonds, angles, improper dihedral angles are constrained with a harmonic potential specific to the mutated residue while the other residues of the protein are restrained¹⁴.

The energy minimization and MD simulations of the structures were performed using the molecular mechanics package CHARMM²¹³, with its 22nd version of the force field²¹⁴. A Particle-Mesh Ewald (PME) method was applied to treat the electrostatic interactions with a dielectric constant of 1 and a cutoff of 10 Å. The Lennard-Jones

potential for any non-bonding interactions were set to 0 beyond 10 Å. The structures were solvated in a water box of TIP3P water molecules²⁶⁴, with the minimum distance of any atom of the protein structure to any face of the box 10 Å or greater. Potassium chloride (KCl) ions were included when the system was solvated to achieve a salt concentration of 150 mM (millimolar), keeping the total charge of the system equal to zero^{114, 212}. Any water molecules that have an oxygen atom closer than 2.6 Å from any of the solute atoms were removed from the system. Bonds involving hydrogen atoms were constrained using SHAKE²¹⁵ during energy minimization and the MD simulation. The MD simulation was performed in an NPT ensemble, where the temperature was maintained to an average of 300 K using a Langevin thermostat and pressure was maintained to 1 atm using the Langevin piston method²⁶⁵.

The solvent of the system was first minimized for 4000 steps using a steepest decent (SD) algorithm while keeping the protein and substrate atoms constrained. In the following steps, a force constant of 1000 kcal/molÅ was applied as a distance-based restraint to the substrate atoms and the atoms within 5 Å of the substrate from the catalytic site. The system was then minimized over 12000 steps using a SD algorithm while the distance-based restraint was scaled down to 800 kcal/molÅ. The solute atoms were then harmonically restrained with a force constant of 700 kcal/molÅ and minimized for 36000 steps using a SD algorithm while the force constant was scaled down to 100 kcal/molÅ.

The system was gradually heated from 100 K to 300 K during 100 ps of the MD simulation with a 1 fs time step with a distance-based restraint for the substrate and active site was applied with a force constant of 150 kcal/molÅ. In the following steps, the atom-based restraint applied to the substrate and active site were removed to begin the MD simulation for the system. The MD simulation of the WT and each of the 68 (12 active and 56 inactive) variants were run for 72 ns. Additionally, most MD simulations equilibrated after 4 ns making a total of 68ns of production for each structure that was run in triplicate for a total of 204 ns production.

The dynamic correlated motions (correlation $C_{(i,j)}$) between each pair of amino acids are calculated with respect to the alpha carbons ($C\alpha$) displacement vectors of the individual amino acid residues as described by equation 1, using the Bio3d ¹²⁰ library in R²⁶⁶⁻²⁶⁸.

$$C_{(i,j)} = \frac{\langle r_i \cdot r_j \rangle}{\langle r_i^2 \rangle^{1/2} \langle r_j^2 \rangle^{1/2}} \quad \text{eq. 3.1}$$

Here, r_i and r_j are the displacement vectors for the $C\alpha$ i and j residues, and the angle brackets denote an ensemble average. A dynamic cross-correlation (DCC) matrix was obtained, using the Bio3d library^{120, 269}, for the WT and each variant consisting of 198² data points each. Every point in the matrix is represented by a real number between 1 and -1,

describing the Pearson correlation related to the movements of each pair of the residues C α within the HIV-1 protease enzyme.

The results from the WT DCC matrix were further evaluated to determine whether residues ($i \neq j$) exhibiting strong negatively (anti) or positively correlated motions, independently, are critical for the function of an enzyme. First the impact from the length of production was evaluated by sequentially removing 10 ns of production from the beginning of the subsequent production and the minimum anticorrelation was compared to each MD production run. In doing so, the residue pair that was found to cause the most anticorrelation was also obtained. The next calculation looked at the negatively correlated motions of the total production by iterating every 0.07 units from -0.10 to maximum anticorrelated value (-0.43). Then the number of unique residues that when substituted result in no activity are divided by the total number of unique residues for each bin, eq. 3.3.

The positively correlated residues were also reviewed by iterating every 0.07 units from 0.10 to maximum correlated value less than 1. Likewise, for each bin the number of unique residues that when substituted result in an inactive variant are divided by the total number of unique residues for corresponding bin. The residue positions between -0.31 and -0.34 were isolated as it was the elbow point found to show an increase in the fraction of inactive variants. The fraction of residue positions between -0.31 and -0.34 were compared to multiple factors, first looking at the impact of the length of production, next

the number of residues known to cause inactivity divided by all unique residues with an experimentally determined activity and last all residues.

The next set of calculations performed used the average known experimentally determined catalytic activity ¹ for the residue within each bin using the same binning methods described for both the positive and anticorrelated residues. Lastly, anticorrelation was split into two groups, group 1) being highly anticorrelated and group 2) being moderately anticorrelated. Lastly, the average experimentally known percent catalytic activity for each group was also calculated.

Machine learning model

Moreover, to predict the protein function using the DCC information implementing the ML (Machine Learning) algorithms, the matrices from the DCC for each of the 56 catalytically inactive (CI) and 12 catalytically active (CA) variant were prepared for analysis using a logistic regression model in R²⁶⁶⁻²⁶⁸ using the glmnet²⁷⁰ library. Logistic regression was used as it is best suited for classifying categorical inputs providing a binary response to uncategorized data. The CA variants are classified as active and numerically encoded as “1”, whereas the CI variants are classified as inactive and numerically encoded as “0”. The total anticorrelation and positive correlations ($i \neq j$) were computed from the DCC of each variant for each of the 99 residues.

The dataset is composed of a 68 x 100 matrix, where 68 represents the 12 CA and 56 CI variants and 100 is the 99 (independent features) residues of a single chain plus 1 (target or dependent feature) from the numerically encoded experimental activity (either 1 or 0 representing either active or inactive variants, respectfully) of HIV-1 protease. The variants were then split and shuffled into training and test sets retaining 20% of the data for testing. Since the data set is small and imbalanced implementing a synthetic oversampling of the minority cases (SMOTE) is highly recommended to improve the distribution of the number of cases between each target which was implemented in R using the `smotefamily`²⁷¹ library²⁶⁶⁻²⁶⁸ providing a better balance of active to inactive cases. Having a greater number of features than the number of cases ends up overfitting and being an unrealistic prediction/model. I used variance inflation factor (VIF) by measuring the variance of the total anticorrelation for each residue pair in order to reduce the number of independent features and the amount of residue pairs with high multicollinearity. The greater the magnitude of VIF found from a pair of residues will likely increase multicollinearity within the dataset and therefore we only keep the residue found to have the greatest correlation to the target “activity”. Such that, if a feature, R1, has a significant correlation with features R3, R5, R8, we only keep the one that has the larger correlation to the target (experimental activity) and remove the rest of features in the dataset. The training dataset was input into a leave-one-out cross-validation (LOOCV), with the `glmnet`²⁷⁰ library in R²⁶⁶⁻²⁶⁸, where the number of folds being equal to the number of cases in the training data set. To evaluate the LOOCV the misclassification error was determined for the datum that was left out.

$$CV(\lambda) = \frac{1}{N} \sum_{i=1}^N L(y_i, \lambda^{(-K(i))}(x_i)) \quad \text{eq. 3.2}$$

Equation 3.2 shown above represents the leave-one-out cross-validation.²⁷² A LOOCV fits a model by is repeated N times for each datum in the training set. Since the classification of active and inactive is binary this method aims to reduce the misclassification rate as well as bias and randomness. The maximum variance for the prediction can be found by using K=N, and therefore, represents the extrema of the data. Both a lasso and ridge penalty²⁷² were evaluated, with similar results however, the lasso penalty was used for the model as it imposes a size constraint on the tuning parameter λ and reduces multicollinearity. The model was then used to evaluate the test set that was evaluated using confusion matrix along with the F1-score, recall, and precision.

Results and Discussion

The symmetrical DCC matrix shown below represents the Pearson correlation of each pair of residues for the WT structure, Figure 3.2. The scale to the right of the Figure 3.2 identifies positively correlated regions (correlation coefficient > 0) are colored from green to red, whereas negatively correlated regions (correlation coefficient < 0) are colored from cyan to dark blue. The DCC plot (on the left side) was depicted for the intra domain (interaction within a chain) as well as the inter domain (interaction between chains). Since this is a homodimer structure, the DCC of the intra domain chain A (region I) and chain B (region II) are similar and for the inter domain between chains A-B (region III) and chains

B-A (region IV) are similar. For simplicity in the analysis, we focus only on the DCC of region I and III. In the inter domain, there is stronger anticorrelation seen between the residues in the inter domain with the rest of the structure; however, there are few residues that show strong positive correlation as well. On the other hand, in the intra domain DCC (region I) there is a good balance of positive and negative correlation between residues; however, looking at the total magnitude of positive vs negative correlation, we see that positive correlation is larger than the anticorrelation for residues i not equal to residue j , despite anticorrelation having a higher frequency (Figure S-3.1).

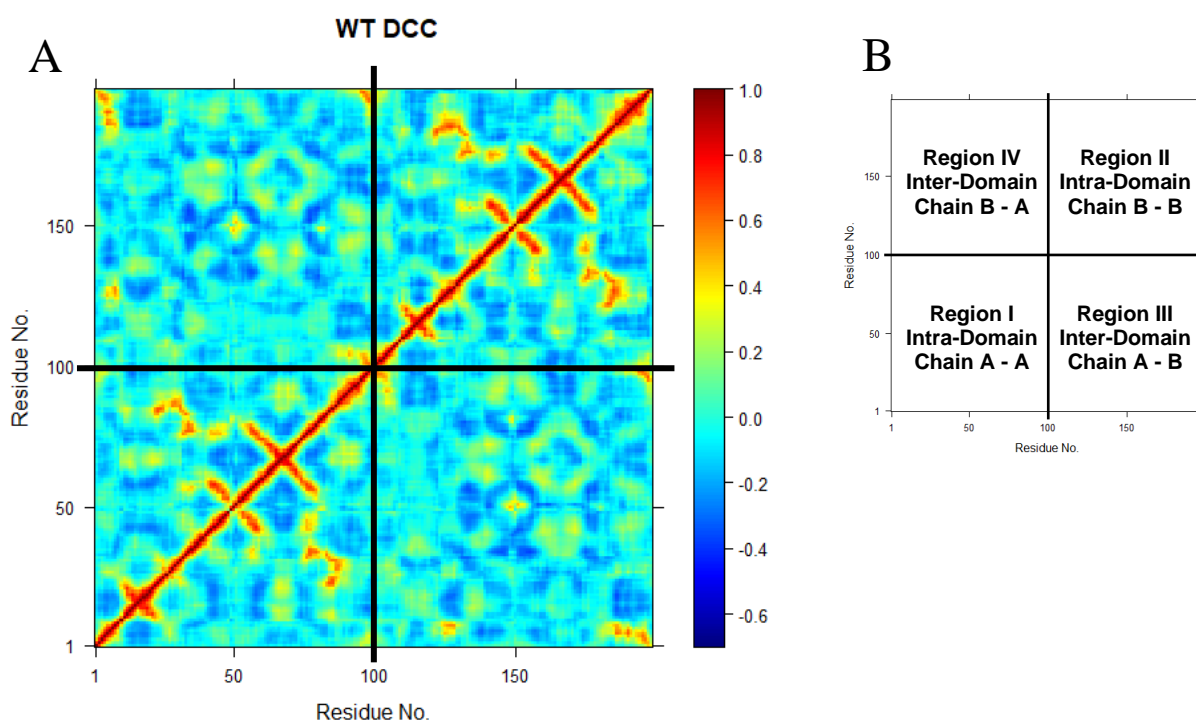


Figure 3.2: (A) Representation Covariance matrix for the motion of C α atoms in WT HIV-1 protease bound to MA/CA substrate. The x and y axis are the residue numbers and the points represent a Pearson correlation value between 1 and -1 based on the average cross correlation value for 3 MD simulations of the WT structure. (B) Shows the regions of intra-protein and domain-domain correlations found in Figure 3.2A.

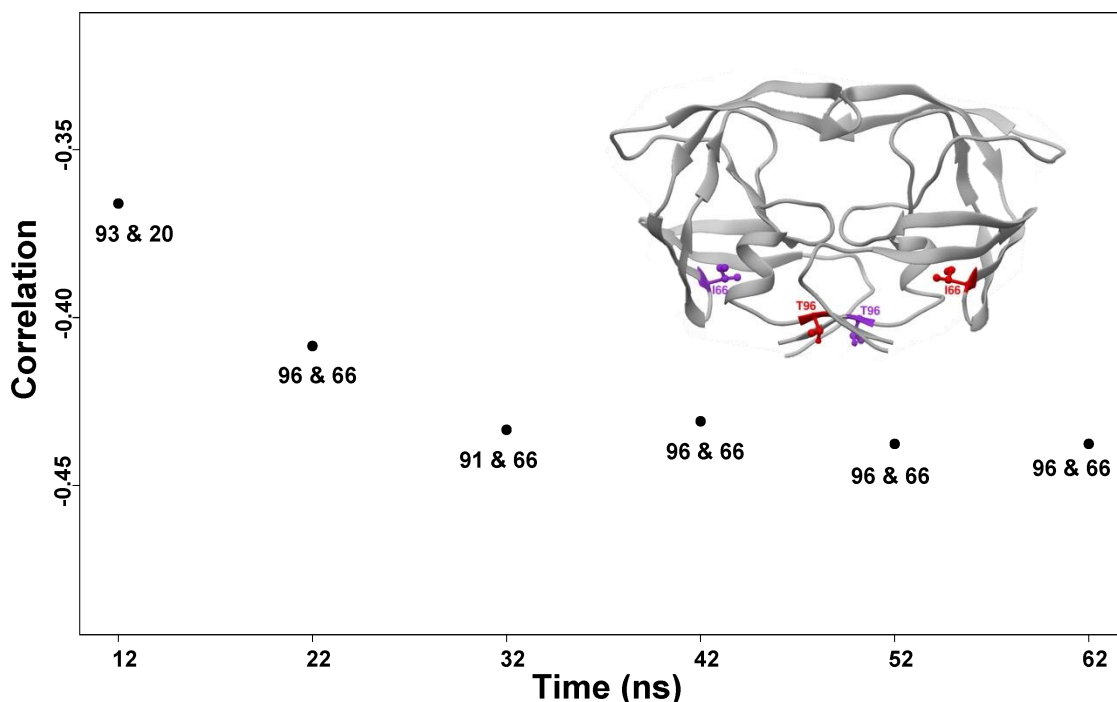


Figure 3.3: Evaluation of the degradation of the minimum anticorrelation observed of the WT structure for every 10 ns of production. The total production of the WT consists of 62 ns and the covariance was calculated for each of the 6 bins above by removing 10 ns from the end of each subsequent production of the WT. The residue pairs that are observed to have the greatest anticorrelation are shown below the respective production.

It can be observed in Figure 3.3 that as the length of the production is reduced the magnitude of anticorrelation is also reduced. It is uncertain whether the magnitude of anticorrelation would increase more as the length of the production is increased further. However, in order to capture the greatest detail that relates dynamics and anticorrelation to the function of an enzyme it is advisable to use more than 32 ns of production run. A major observation that can be found in Figure 3.3, is that, as the length of production is reduced, we can see the residue pairs that have the greatest anticorrelation also change. Additionally, the residues found to exhibit the greatest anticorrelation are residue pairs found in separate

domains. Further scrutiny suggests that since the residue pairs change as the length of the production is reduced that the moderate correlations would also change as the production is reduced as well. Also, another factor that indicates the stability from the length of production is seen by having the same residue pairs are presented as the production length increases beyond 32 ns. Performing the same calculation with respect to the maximum positively correlated residues (greater than 0 and less than 1) does not show any pattern as we observed for the case of anticorrelation, Figure S-3.1. Secondly, the maximum positively correlated residue are more often neighboring residues, in the same chain.

The anticorrelation was segmented into bins using the DCC matrix from the WT structure for the total length of production (72ns). We started binning the anticorrelation from -0.10 to the maximum identified anticorrelation of -0.43 to examined if the magnitude of anticorrelation for certain residues can be attributed to a protein's catalytic activity, Figure 3.4. All the number of residues within each anticorrelated region are segmented every 0.07 units making 12 bins along the x-axis and the fraction of the number of residues known to cause a CI variant to the total anticorrelated residues, eq 3.3, are evaluated along the y-axis.

$$y = \frac{n_{CIV}}{n_{Total}} \quad \text{eq. 3.3}$$

Here n represents the number of residues, n_{Total} and n_{CIV} represent the total number of anticorrelated residues and the number of residues known to cause a CI variant seen in the n_{Total} for the corresponding bin, respectively. The following observations are

seen in Figures 3.4 & 3.5. First, at the lower magnitude of anti/positive correlation we can see most residues are present and likewise for the number of residues found to be associated with a CI or CA variant. In other words, as the number of residues having a larger magnitude of anti/positive correlation decreases the number of residues known to cause a CI variant shows an increase. The first observation can lead us to the second point. It can be indicative that if the residues show higher anticorrelation with the other residues, the probability that variation on those residues causes inactivity would be greater. In other words, though the proportion of inactive variants does not increase significantly as the anticorrelation increases in magnitude the residues that result in an inactive variant can influence the dynamics of anticorrelated motion.

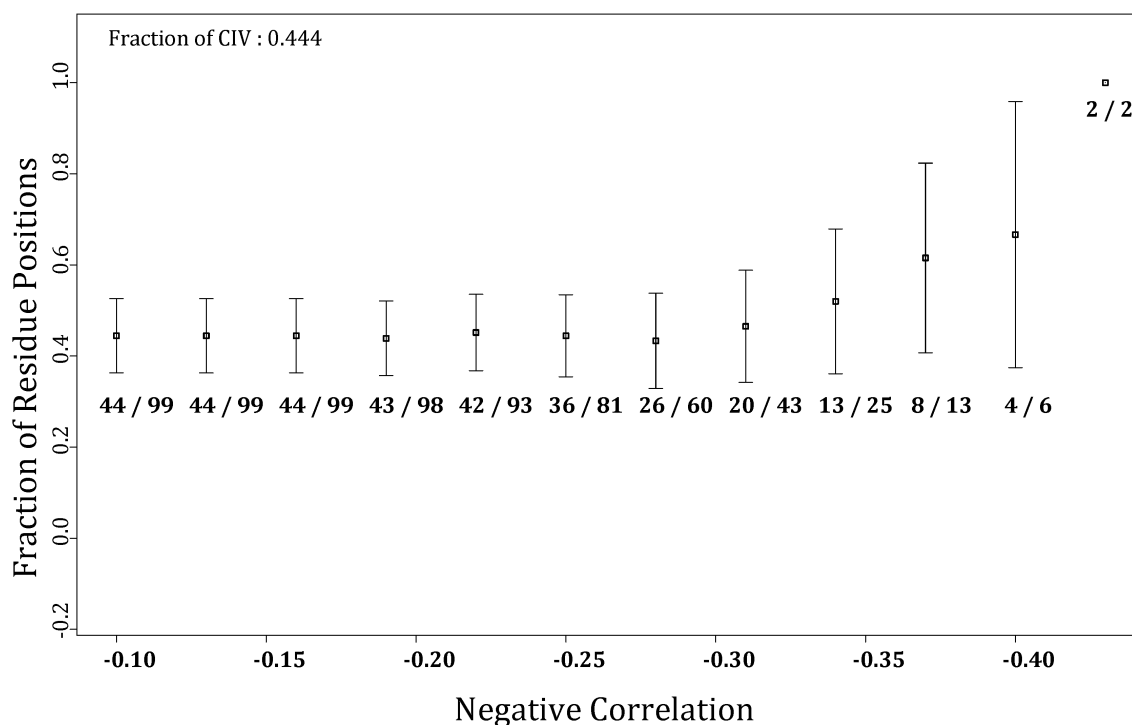


Figure 3.4: This figure examines the relationship between the degree of negatively correlated motion associated with each unique residue position in HIV-1 protease and the impact of a substitution resulting in an inactive variant toward the MA/CA substrate. If anticorrelated motion are strongly linked to enzyme activity in this system, one might expect to observe a positive slope in this figure suggesting that specific sequence positions exhibiting stronger correlated motions would also be more likely to eliminate catalytic activity upon substitution. The data here demonstrates, though there is not a positive slope, that the substitutions that result in an inactive enzyme strongly influence anticorrelation. The number of residues in each chain were collected within an iteration of every 0.03 units from -0.18 to the maximum negative correlation (-0.43). Then, the sequence positions that when mutated result in the protein having no noticeable catalytic activity (44 unique residue positions in each chain of the homodimer) are divided by all residue positions (99 residue positions in each chain of the homodimer) making up the fraction of catalytically inactive (CI) residue positions. The standard error was calculated from the confidence interval of the proportion, $z_{0.95} \sqrt{(p(1 - p))/n_p}$, shown by the error bars.

Similarly, we examined whether the residues exhibiting positively correlated motions within the DCC of the WT structure could be linked to the catalytic activity of the HIV protease, Figure 3.5, the values of positive correlation are segmented into 12 bins along the x-axis and the fraction of the number of residues known to cause a CI variant to the total positive correlated residues, eq 3, is evaluated along the y-axis, Figure 3.5. As we stated in the section above (Observation for Fig 3a), we see that the higher number of residues caused CI variant occurs in the lower magnitude of positive correlation and as the magnitude of positive correlation increases the less variants caused CI variant fell in this range of residues. However, there is a difference between highly positive correlated residues vs highly negative ones. In the larger positive correlation range, the probability of having variation lead to CI variant is not as high as for the larger negative correlation range. For instance, we have the following probabilities for the last three bins of positives correlation vs anticorrelation: 0.46, 0.37, 0.5 vs 0.61, 0.67, 1.00. This can show how

important the influence of anticorrelation from the CI residues or in other words, in the relationship of protein function and activity.

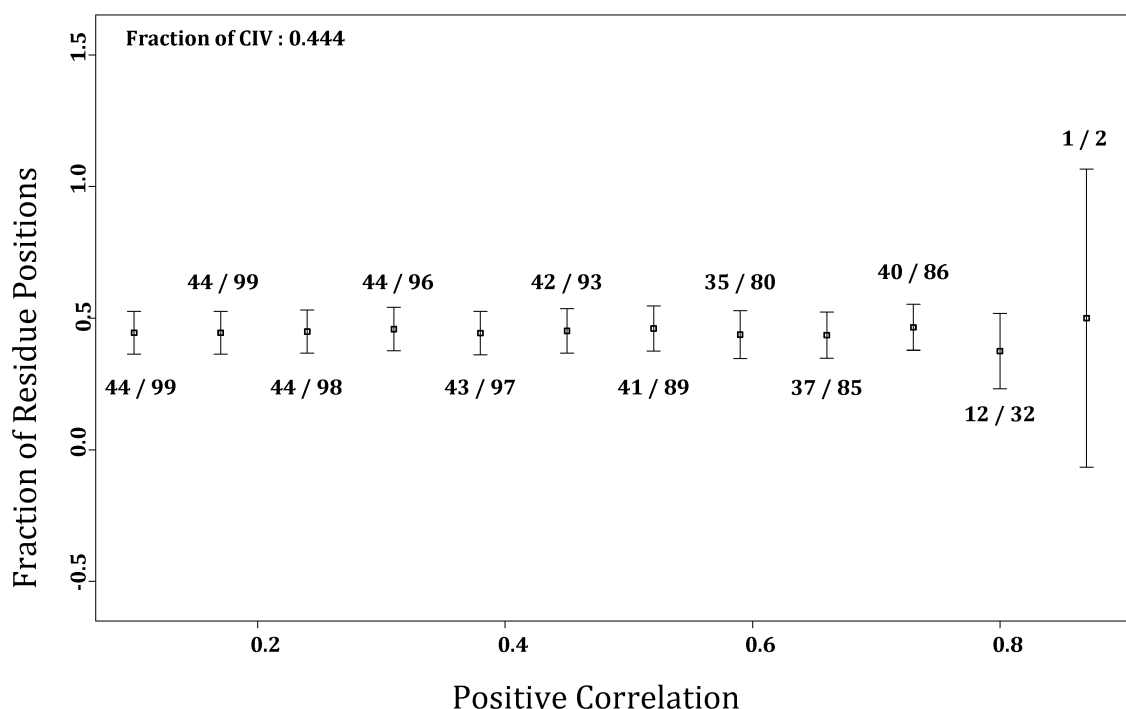


Figure 3.5: This figure examines the relationship between the degree of positively correlated motion associated with specific sequence positions in HIV-1 protease and the impact of a substitution of those residue positions on the peptidase activity toward the MA/CA substrate. If correlated motion were strongly linked to enzyme activity in this system, one might expect to observe a positive slope in this figure suggesting that specific sequence positions exhibiting stronger correlated motions would also be more likely to eliminate catalytic activity upon substitution. The data here demonstrates significant sensitivity of the degree of correlated motion with respect to the impact of substitution on enzyme catalytic activity. The number of unique residues were collected by iterating of every 0.07 units from 0.10 to maximum correlated motion not equal to 1. Within each bin the residue positions known to result in the protein having no noticeable catalytic activity (44 residue positions in each chain of the homodimer) are divided by all residue positions identified as having correlation in that range, making up the fraction of residue positions. The standard error was calculated from the confidence interval of the proportion, $z_{0.95} \sqrt{(p(1 - p))/n_p}$, shown by the error bars.

In the next two figures the correlation values were segmented using the same binning method that was applied for Figure 3.5 A&B. However, the average known catalytic activity was calculated from the experimental data¹, whereas a bin consisted of only residues with known experimental activity and the average of the bin was then calculated for a bin, and is represented along the y-axis, Figure 3.6A. We expect that the dynamics of the enzyme is anticipated to impact the average catalytic activity of the residues and there should be a decrease in the average catalytic activity as the absolute value of the magnitude of anticorrelation increases. Figure 3.6A shows that the average activity of residues decreases as the absolute value of the magnitude of anticorrelation increases. In other words, it indicates that there is a relationship between anticorrelated residues and the activity of an enzyme. Moreover, it shows that when the residues are anticorrelated but not strongly, the variation on these residues can lead to the activity of enzyme. Similarly, in Figure 3.6B, if the average catalytic activity of residue positions has greater influence on positive correlation is considered to affect the dynamics of the enzyme more there should be a decrease in the average catalytic activity as the positive correlation increases. These results from Figure 3.6B suggest that residues exhibiting significant positively correlated motions are not as critical for the function of the WT structure of HIV-1 protease.

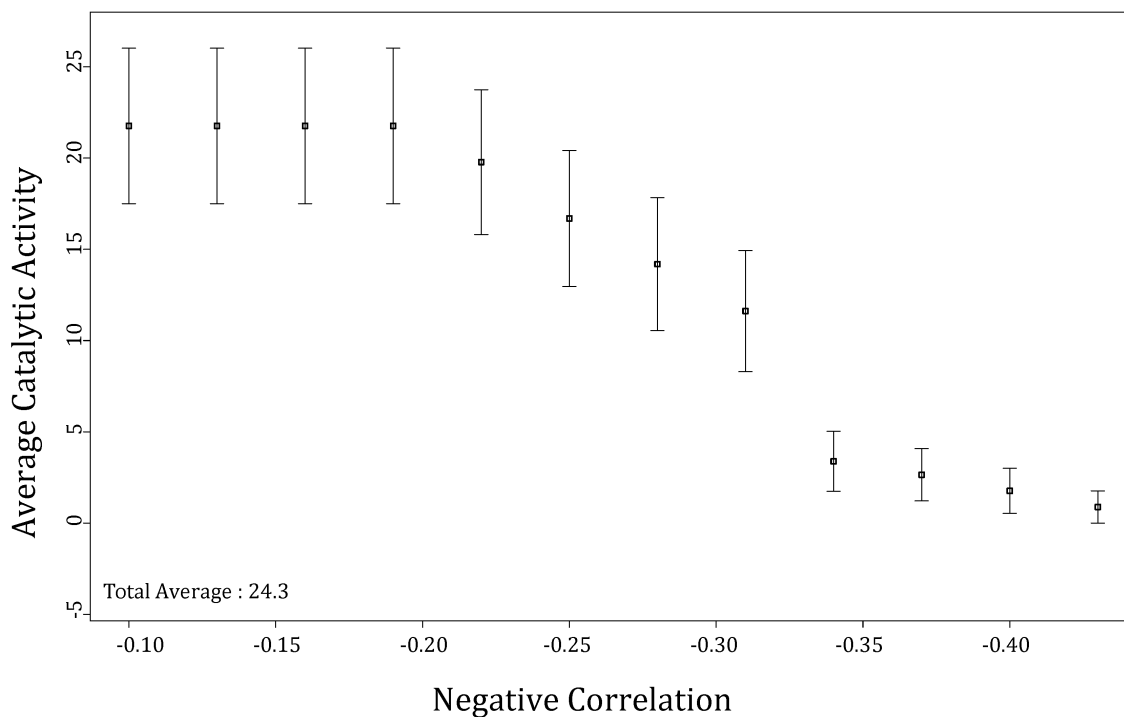


Figure 3.6A: This figure examines the relationship between the degree of negatively correlated motion associated with specific sequence positions in HIV-1 protease and the impact of the average peptidase activity toward the MA/CA substrate. If correlated motion were strongly linked to enzyme activity in this system, it can be expected to observe a negative slope in this figure suggesting that specific sequence positions exhibiting stronger anticorrelated motions would also be more likely to eliminate catalytic activity upon substitution. The data here demonstrates that the degree of anticorrelated motion with respect to the impact of substitution on enzyme catalytic activity is statistically significant. The number of residues in each chain were collected within an iteration of every 0.03 units from -0.1 to the maximum negative correlation (-0.43). Then, the average of the known catalytic activity was calculated for the sequence positions within each bin. The standard error was calculated using $= \sqrt{\frac{(x-\mu)^2}{N}}/\sqrt{N}$, shown by the error bars.

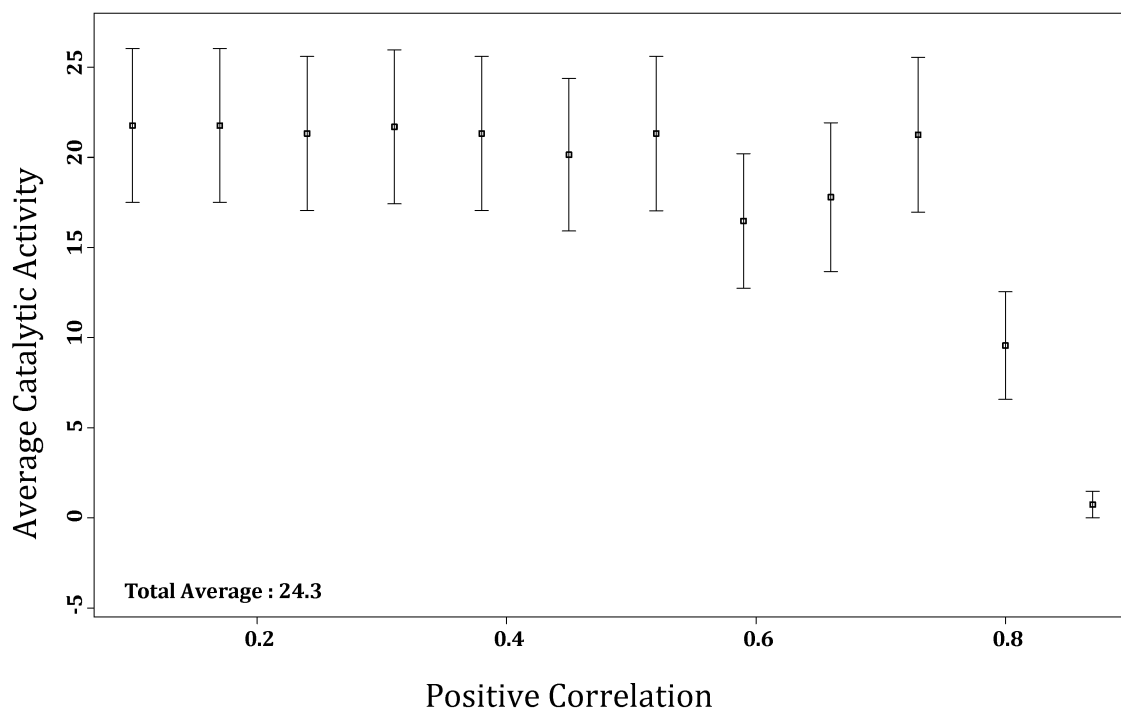


Figure 3.6B: This figure examines the relationship between the degree of positively correlated motion associated with specific sequence positions in HIV-1 protease and the impact of the average peptidase activity toward the MA/CA substrate. If correlated motion were strongly linked to enzyme activity in this system, it would be expected to observe a negative slope in this figure suggesting that specific sequence positions exhibiting stronger correlated motions would also be more likely to eliminate catalytic activity upon substitution. The number of residues in each chain were collected within an iteration of every 0.07 units from 0.10 to the maximum correlation less than 1. The average of the known catalytic activity was calculated for the sequence positions within each bin. The standard error was calculated using $SE = \sqrt{\frac{(x-\mu)^2}{N}}/\sqrt{N}$, shown by the error bars.

The residues, based on the WT DCC matrix, were separated into two groups, Table 3.1. Group 1 represents residues that are considered to be highly anticorrelated ($C_{ij} \leq -0.31$) and Group 2 are residues that are found to be moderately anticorrelated ($-0.10 > C_{ij} > -0.31$), Table 3.1. The objective was to compare the effect of substitutions within each group of these residues on the activity of the enzyme. The residues exhibiting the strongest

negatively correlated motions should exhibit the largest changes in the activity upon substitution since anticorrelation is hypothesized to be critical for the function of the enzyme. The results seen in Table 3.1 indicate that almost half of the residue in either group are reported as having no catalytic activity compared to the WT structure. In other words, the amino acid positions that show greater negatively correlated motions are slightly more sensitive towards substitutions in comparison to positions that show weak negatively correlated motion. Likewise, the average percent catalytic activity within each group is ~4% and ~20% for Group 1 and 2 respectively, showing the slight sensitive can dramatically reduce the average activity, Table 3.1.

Table 3.1: Group 1 corresponds to the positions showing strongly negative correlation and Group 2 corresponds to the positions showing relatively weak negative correlation.

	Correlation coefficient (C_{ij})	Number of total residue positions in each group	Fraction of positions that upon substitution result in an inactive variant	Average % activity of the substitutions of each group
Group 1	$C_{ij} \leq -0.31$	47	0.53	4.06
Group 2	$-0.10 > C_{ij} > -0.31$	99	0.44	19.67

Though it is observable that the CI residues play a critical role in influencing the dynamic anticorrelated motions they only make up a little more than half the residue positions that are observed in group 1. The proportion of CI residue is smaller in group 2 since every residue is found to exhibit anticorrelated motions in this group. The last column shows that, though active variants can influence the average activity in either group, the inactive variants have a greater influence on anticorrelated motion.

In further analysis, we perform ML algorithm to capture the relation between the DCC (positive/negative correlation) of the residues and the protein function/activity and whether we can predict if a single point substitution will lead to inactive/activity of the protein using the DCC map from the MD simulation. Before performing the ML algorithm, we need to take into account several factors in the active and inactive data sets (active/inactive datasets obtained using DCC for the CA/CI variants). The dataset is composed of 68 data points with each row belonging to one of the 12 CA and 56 CI variants. There are 99 columns (independent features) representing the residues in a single chain of HIV-1 protease and the value of each of these columns is the total anticorrelation rate of the corresponding residue with other residues (except with itself). Based on our observation in the previous sections, we collected only the total anticorrelation of residues. Also, we have a binary classification problem with the target value of 1 for the CA variants and 0 for the CI variants.

Next, the dataset (shuffled) was split (randomly) into training and testing sets with 80 and 20 percent of datapoints in each, respectively. To remove the potential overfitting issue from having a large number of features in our somewhat small database (58 CI and 12 CA variants) we take advantage of the multicollinearity concern where there is a significant dependency between these independent features. By removing the multicollinearity issue, it leads us to a smaller set of 12 independent features, Figure S-3.2. In order to further reduce the potential bias in the system due to imbalanced dataset (58 total variants), synthetic minority oversampling (SMOTE) was implemented on the

training data only to increase the number of active cases from 12 to 42. Using SMOTE allows for a less bias distribution of different classes in the dataset. The training data was fit to the logistic regression using a leave one out cross validation (LOOCV) technique. The LOOCV is used to complement the SMOTE and feature reduction techniques to adequately reduce bias and can be a good choice due to our small dataset. The LOOCV was applied using a binomial algorithm with a lasso penalty, the same calculation was done with a ridge penalty that provided similar results. A LOOCV can be used to find a minimum lambda that can be applied to the prediction algorithm as the normalization factor for performing the logistic regression calculation on the unclassified data. The misclassification error with the $\log(\text{Lambda})$ on the x-axis and misclassification error on the y-axis, the vertical line representing the best lambda from the cross validation (Figure S-3.3). The fitted model was then validated on the training set and used to evaluate the performance of the model on the unseen datapoints from the test set.

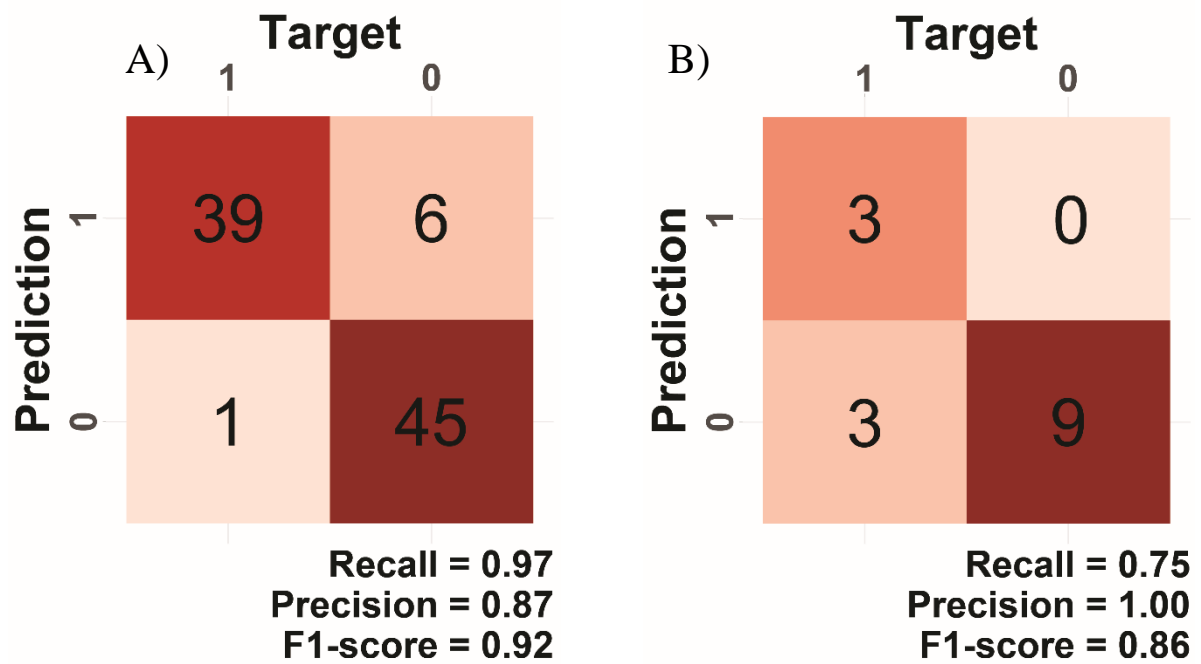


Figure 3.7: Confusion matrices for the XGBoost model performance **A)** is the performance result of the training set and **B)** is the performance result of the test set with the recall, precision and F1-score shown at the bottom right.

The performance of the trained model is shown in Figure 3.7 for the test set along with other metrics such as F1-score, recall and precision. The result shows that the model successfully can predict 9 CI variants out of 12 CI variants and 3 CA variants out of 3 CA variants. The misclassification rate for the test set is 0.20 that belongs to 3 false positives (FP). However, the overall performance of the model provides a good prediction of both classes, specifically minority (active) cases.

Table 3.2: Evaluation of 8 runs of the XGBoost model using different random seeds.

	Train F1-score	Test F1-score	Train Recall	Test Recall
Set 1	0.955	0.857	0.977	0.75
Set 2	0.947	0.857	0.900	0.75
Set 3	0.918	0.857	0.975	0.75
Set 4	0.938	0.857	0.882	0.75
Set 5	0.918	0.857	0.849	0.75
Set 6	0.918	0.857	0.849	0.75
Set 7	0.918	0.857	0.849	0.75
Set 8	0.988	0.889	0.976	0.80

To split the dataset into training and test sets, we implemented “seed” parameter which leads to slightly different resampled data for the training and test sets with a different value of the seed. In order to evaluate how different values of the seed impact the model and its performance, we used 8 different values for the seed in Table 3.2. This result indicates that the results are reproducible using different random seeds.

Conclusions

In this study, we examine whether the residue positions found to possess significant correlations have any influence on the experimental determined activity¹ (variant activities). Though the inference from the present study shows a clear trend between activity and correlated dynamics for HIV-1 protease; in general, different proteins may

require different parameters, either longer or shorter MD simulations or reduced representations, to assess the impact correlated dynamics has on the activity. Whether the CA or CI variant can also be found active or inactive with any of the other substrates common to HIV-1 protease is also unknown. Though the system has equilibrated after a few ns, it is certain that observing the low frequency modes that account for the global anharmonic frequency could require greater than 32 ns of production. As a good measure, it is also evident that running simulations less the 32 ns dramatically reduces the number of cases observed to influence anticorrelation.

The study of such progressively robust enzymes, such as HIV-1 protease, demonstrates that the dynamics plays a significant role as a constraint on its evolution. It is evident that as the magnitude of anticorrelation increases, becomes more negative, though less residues are observed more often they are associated with residue position that results in a CI variant. The characterization of the correlated motion from the residue positions of HIV-1 protease can allow scientists to have a greater understanding of how the dynamics plays a role in the function of an enzyme. Evaluating the residue positions from the DCC matrix of the WT structure and variants in this study provides evidence that dynamics can be a significant constraint on the evolution of HIV-1 protease. Examination of the obscure dynamic changes between a C2 symmetrical dimer and its variants showed that the greatest anticorrelated motions occur within the inter-domain region. Since these anticorrelated motions provide evidence of long-range inter-domain signaling that can aide in developing better classification of protein-protein interactions and how they can modify

an enzyme's function. HIV-1 protease's robustness towards substitutions and promiscuity towards several substrates makes it a very difficult system to characterize. However, from the thorough investigation conducted here we are able to identify the factors that influence change in the system and provided significant evidence that correlated motion acts as a constraint on the activity of HIV-1 protease. The statistical methods applied here aided us in linking the dominate correlated motions from each pair of residue positions to the activity of the variants. The machine learning (ML) algorithms used to solve the binary-classification with regards to the function of the variants are able to predict the activity of an unknown variant effectively, based on the correlated dynamics. Furthermore, the methods applied here offer and can be applied to gain a better understanding of mechanisms associated with allosteric interactions.

CHAPTER IV

UNDERSTANDING THE ROLE OF DYNAMICS USING MUTUAL INFORMATION AND DEVELOPING A ROBUST MACHINE LEARNING MODEL TO PREDICT THE FUNCTION OF HIV-1 PROTEASE VARIANTS

Introduction

Variant structures arise due to the improper transcription of the amino acid sequence causing an insertion, deletion, or substitution of one or more of the amino acids commonly observed in the wild-type (WT) structure. In this chapter we focus specifically on substitutions and how the dynamics is used to regulate the function associated of these variant structures. Experimentalists and theorists use substitutions to examine the causality which is the general nature and function that represents the WT structure. In the case of a single point substitution, minor or major adjustments to the protein's structure and function can drive over or under expression of a variant proteins caused by signal and response mechanisms that can alter the expected output that is required by the system^{155, 273-276}. The unique chemical and physical characteristics of any amino acid imposes an exclusive local environment when replaced by residue^{33, 150, 277, 278}. The amino acids that make up the WT

structure are assigned to provide optimal performance towards consuming its intended substrate and the amount of product produced. Though some of the variants can supply faster or more product, they may produce more product than necessary or consume the substrate faster than it comes available making them less favorable^{33, 150, 277, 278}. Effectively, the inherent conformational states provided by the WT amino acid sequence is a means to regulate the correct amount of product is available and the rate of consumption is acceptable^{249, 279-281}. To understand more of how a protein's structure and function can be related, we need to examine how a residue is able to distinguish changes from information provided at distant sites that alter the function of the structure.

The alterations caused by a substitution or some other event at a location on an enzyme can regulate the function of another non-adjointing distant location of the enzyme, which is known as allostery⁴⁴⁻⁵⁰. The study of protein communication networks, regulation, and stability can provide scientists the methods to design and understand drug delivery, disease, and cellular processes better⁵⁰⁻⁵⁴. There are several approaches considered to aid in interpreting how the allosteric affects propagate, how a change at "A" can cause an affect at "Z". The processes within a protein are highly important for regulating the activity of an enzyme in order to preserve cellular function⁵⁰⁻⁵⁴.

Though the changes are not always directly initiated within the catalytic site, the residues within the catalytic site of an enzyme have more responsible than most and therefore changes to those positions can have larger consequences^{163, 282-285}. It is also

important to note that the residues within a protein-protein interface, for oligomeric proteins, have a high degree of importance as they can disrupt the oligomerization of a protein^{163, 282-285}. The study of amino acid substitutions can advance the understanding of evolutionary pathways and the drugability of an enzyme^{96, 274}. Additionally, investigating the detailed characteristics of the contributing amino acids along with the functional properties of the residue position can be used to better understand the allosteric mechanisms^{108, 239, 240}.

Biophysicists and biochemists have adopted a statistical method used to improve signal processing of communication devices called Mutual Information (MI)¹³¹⁻¹³³. MI is a statistic method within Information Theory that focuses on the communication of data sets^{131, 132, 134-136}. The theory behind MI was developed by Shannon in 1948, defining what is termed Shannon entropy (eq. 4.1): is the non-linear measure of the rate of information associated with the variables being studied, similar to the theory of entropy from thermodynamics^{131, 132}. In Shannon entropy, as the choice becomes certain the entropy vanishes, goes to 0, meaning information increases as the uncertainty of the system/observation increases^{131, 132}. When considering the joint probability distribution, the concept of uncertainty remains the same as that of Shannon entropy, whereas, knowing less about the information of residues i and/or j will increase the joint Shannon entropy^{67, 134-137}. Many have adopted this theory to understand the uncertainty and the noise associated with a biological, chemical, or physical phenomena^{67, 134-137}. Thus, MI can be used to explore how the linear and non-linear mutual dependency between distant residue

interactions and how they can be used to interpret evidence of allosteric communication¹³¹⁻¹³³.

Consequently, much of the current studies utilize modified states or methods that require further amplification of the signal observed from the MI calculation^{67, 134-137}. Ultimately a deeper understanding of the topological features that influence a proteins function caused by information sent from distant residues should be evaluated further^{67, 134-137}. Within a protein structure each amino acid residue experiences many different environmental conditions and interactions such that: surface residues mostly interact with the solvent, bulk and interface residues primarily associate with other residues, and the binding pocket residues come in contact with the solute or solvent, if accessible^{93, 141-145}. Consequently, the forces and interactions imposed on each residue can vary significantly, depending on the location and neighboring residues, constituting a considerable factor for the identity and/or pathway of allosteric communication^{23, 50, 76, 146}.

In order to develop a clear understanding of how information is passed from one residue to the next we focus on a symmetric robust protease, the human immunodeficiency virus type 1 (HIV-1) protease. Besides playing an important role in the maturation process and viral replication of the immature AIDS virus, HIV-1 protease is well characterized and common analog exploited for drug efficacy^{148, 182, 286-289}. The inefficient replication process of HIV-1 protease allows for the competition of functional variants that makes it challenging to find adequate protease inhibitors¹⁹¹⁻¹⁹³. The competition of functional

variants combined with the responsible for the cleavage of Gag and Gag-Pol poly proteins at eleven sites, HIV-1 protease has a diverse selection of substrates^{148, 182, 286-289}. In the presence of a protease inhibitor drug, functional variants can cause resistance against the drug binding to the protease making them preferentially selected for viral replication¹⁹¹⁻¹⁹³. This process of selecting different functional variants of the HIV-1 protease can also be thought as enzymatic evolution.

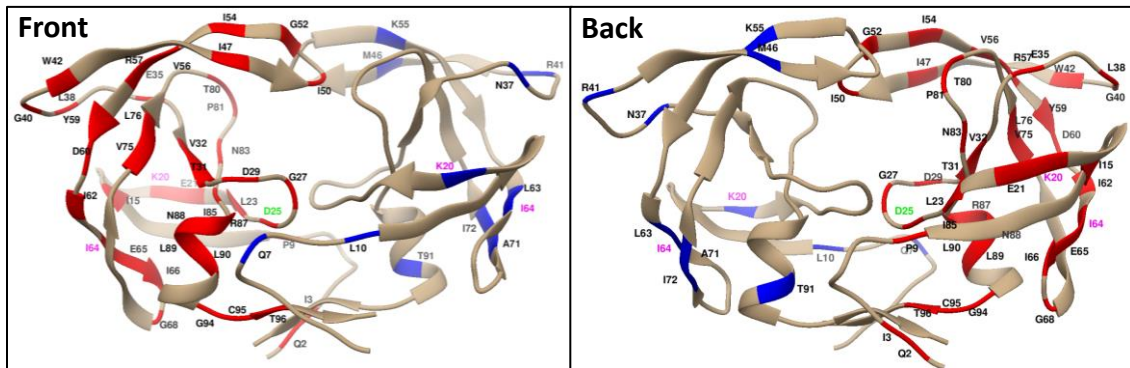


Figure 4.1: The figure shown above is a front and back view of the HIV-1 protease (PDB id: 1hwx). Substitutions were carried out on both chains but the location of active and inactive substitutions are shown independently on a single chain. Each residue position known to be as active or have greater activity than the WT are highlighted in green and those shown to have zero activity compared to the WT are shown in red. Of the 99 residues that make up each chain, there are 42 unique positions that are found to cause an inactive variant and 12 unique positions that are found to cause a variant with activity equal or greater than the WT structure.

The HIV-1 protease is a homodimer of two identical 99 amino acid chains for a total of 198 residues. Besides the catalytic site, HIV-1 has several functional regions that support the operation of the robust C2 symmetrical homodimer that include the flap, elbow, hinge, core, and the protein-protein interface^{54, 186-190}. Symmetry has been observed to

contribute to the evolutionary stability and cooperative interactions between the subunits of many oligomeric proteins, like that of HIV-1 protease^{54, 186-190}.

Here we utilize the publication of an in-vitro study of the experimental catalytic activity where 107 HIV-1 protease variants activity was measured towards the cleavage of the matrix-capsid (MA/CA) polypeptide domain¹. The experimental data¹ demonstrated that a single point substitution that occurs in both chains of the enzyme can diversely effect the catalytic activity of HIV-1 protease towards the cleavage of MA/CA. Closer inspection of the data¹ presents 68 variants of the HIV-1 protease that can be considered highly active (12) or zero activity (56) with respect to the WT structure, Figure 4.1. Modeling the HIV-1 protease (WT) and variant structures in the bound state with one of its natural substrates can offer insight of a bridge between the dynamics and the function of robust symmetrical enzymes. The information provide by Martinez et.al.¹ is used in conjunction with the recently determined structure of HIV-1 protease in bound state with the MA/CA peptide substrate²⁰⁸. Studies such as Martinez et.al.¹ as well as highly distinguished protein structures can assist researchers to explore the relevance of the protein dynamics and the role of allosteric signaling plays in the enzymatic evolution. Probing the dynamic interactions of the residues of an enzyme and its variants can offer insight for the developmental importance of allosteric signaling and their connection to a variant's experimental activity.

Here we seek to understand how the non-linear dependency identified from the interactions of each pair of residue positions, using Mutual Information (MI), can be used to describe the dynamics of HIV-1 protease. Residues that receive more or less information than what is commonly observed from the WT may be associated with allosteric “hotspot” and/or disrupt the allosteric pathway/s within a protein. Interference to the allosteric pathway and/or an allosteric “hotspot” can adversely impact the function of the protein that can then be exploited as a target during drug trials. In this study we look to uncover if the communication sent or received by a substituted residue or the residue’s they interact with most can act as a constraint on the function of HIV-1 protease. We also use machine learning (ML) algorithms suited for solving binary-classification problems in order to assess if there is a connection between the non-linear pairwise interactions of HIV-1 protease variants with their experimentally determined activity. Furthermore, using the ML algorithms to predict the impact caused by a single point substitution can aid scientists in better understanding the allosteric mechanisms that can influence the activity of a protein.

Methods

The models of the WT and variant complexes of HIV-1 protease with the MA/CA substrate were prepared using the following methodology. The structure of the MA/CA substrate and a tetra coordinated water molecule were extracted from a PDB file (PDB id: 1KJ4)²⁰⁸ of HIV-1 protease. The MA/CA substrate and tetra coordinated water molecule were placed into the binding pocket of the WT structure (PDB id: 1HXW)²⁰⁹, with the same

amino acid sequence as the WT structure²¹⁰. The substrate and tetra-coordinated water were transferred to the WT structure by aligning the backbone atoms of the active site (within 4.5 Å of ligand) for both PDB structures. Using software called MODELLER¹⁴ each variant was created such that a single point substitution was made in both chains of an appended WT structure. MODELLER¹⁴ allows the use of CHARMM²¹³ packages, such that, the harmonic potentials specific to the bonds, angles, improper dihedral angles of the substituted residue are constrained while the other residues of the protein are restrained.

The structures were solvated in a water box of TIP3P water molecules²⁶⁴, with the minimum distance of any atom of the protein structure to any face of the box 10 Å or greater. Potassium chloride (KCl) ions were included when the system was solvated to achieve a salt concentration of 150 mM (millimolar), keeping the total charge of the system equal to zero^{114, 212}. Any water molecules that have an oxygen atom closer than 2.6 Å from the protein and substrate were removed from the system. The energy minimization and the MD simulations of the structures were performed using the molecular mechanics package CHARMM²¹³, with its 22nd version of the force field²¹⁴. A Particle-Mesh Ewald (PME) method was applied to treat the electrostatic interactions with a dielectric constant of 1 and a cutoff of 10 Å. The Lennard-Jones potential for any non-bonding interactions were set to 0 beyond 10 Å. Bonds involving hydrogen atoms were constrained using SHAKE²¹⁵ during the energy minimization and the MD simulation. The MD simulation was performed in an NPT (constant pressure and temperature) ensemble, where the temperature was maintained

to an average of 300 K using a Langevin thermostat while the pressure was maintained to 1 atm using the Langevin piston method²⁶⁵.

The solvent of the system was first minimized for 4000 steps using a steepest decent (SD) algorithm while keeping protein and substrate atoms constrained. In the following steps, a force constant of 1000 kcal/molÅ was applied as a distance-based restraint to the substrate atoms and the atoms within 5Å of the substrate from the catalytic site. The system was then minimized over 12000 steps using a SD algorithm while the distance-based restraint was scaled down to 800 kcal/molÅ. The solute atoms were then harmonically restrained with a force constant of 700 kcal/molÅ and minimized for 36000 steps using a SD algorithm while the forces constant was scaled down to 100 kcal/molÅ.

The system was gradually heated from 100 K to 300 K during 100 ps of MD simulation with a 1 fs time step with a distance-based restraint for the substrate and active site was applied with a force constant of 150 kcal/molÅ. In the following steps, the atom-based restraint applied to the substrate and active site were removed to begin the MD simulation for the system. The MD simulation of the WT and each of the 68 (12 active and 56 inactive) variants were run for 72 ns. Additionally, most MD simulations equilibrated after 4ns making a total of 68ns of production for each structure that was run in triplicate for a total of 204 ns production.

The alpha carbon (C α) coordinates of the structure's trajectory were extracted using the Bio3d¹²⁰ library with the R programming language²⁶⁶⁻²⁶⁸. The C α coordinates for each residue were subsequently scaled by calculating the magnitude of the vector.

$$|a| = \sqrt{x^2 + y^2 + z^2} \quad \text{eq. 4.1}$$

Next, the residues along with their scaled C α coordinates were input into a K-means algorithm (implemented the caret²⁹⁴ library with R²⁶⁷⁻²⁶⁹) to cluster residues based on their scaled C α coordinates for all cases over the length of the MD production²⁹⁰⁻²⁹². An elbow method was applied to determine the optimal number of bins used to discretize the residues based on the scaled C α coordinates, see Figure S-4.1. The C α coordinates were then numerically encoding using the discretize function from the infotheo²⁹⁴ library with R²⁶⁶⁻²⁶⁸ and the 115 bins. The MI was then calculated using the infotheo²⁹⁴ library with R²⁶⁶⁻²⁶⁸. The MI for the scaled C α coordinates of each residue pair was calculated generating a square matrix of 198² data points. The total MI of each residue was calculated from the pairwise MI by focusing exclusively on the interactions within a single domain, though the intra-domain MI for each residue was also calculated.

Shannon entropy^{131, 132} describes the probability mass function that r_i or r_j share some measurable quantitative value, equation 4.2. In other terms, it is the individual entropy of one variable and the conditional entropy of another variable^{131, 132}. MI is the relative entropy between the joint distribution and the product distribution, eq. 4.2. The

joint Mutual Information is the reduction in the uncertainty of one variable due to the knowledge of another.

$$H(R_i; R_j) = - \sum_{(r_i \in R_i)} \sum_{(r_j \in R_j)} p(r_i, r_j) \log p(r_i, r_j) \quad \text{eq. 4.2}$$

$$I(R_i; R_j) = \sum_{(r_i \in R_i)} \sum_{(r_j \in R_j)} p(r_i, r_j) \log \frac{p(r_i, r_j)}{p(r_i)p(r_j)} \quad \text{eq. 4.3}$$

The total MI of each residue was calculated from the pairwise MI using the infotheo²⁹⁴ library in R²⁶⁶⁻²⁶⁸ for each variant and the WT, focusing mainly on the MI within a single domain, though the total MI for the inter-domain (Please see Figure 4.3A) interactions of each residue was also computed separately. Next, the total MI for the active and inactive variant cases were separated into variant residues with low, high, and neutral MI compared to the WT. Residues with low MI are found if 45% or less of the variants for a residue are shown to have less MI than the WT, Figure 4.2. Residues with high MI are found if 55% or more of the variants for a residue are shown to have more MI than the WT, Figure 4.2. Whereas a residue was found to be neutral if in the variant cases it shows a total MI of more than 45% and less than 55% of that in the WT structure, Figure 4.2.

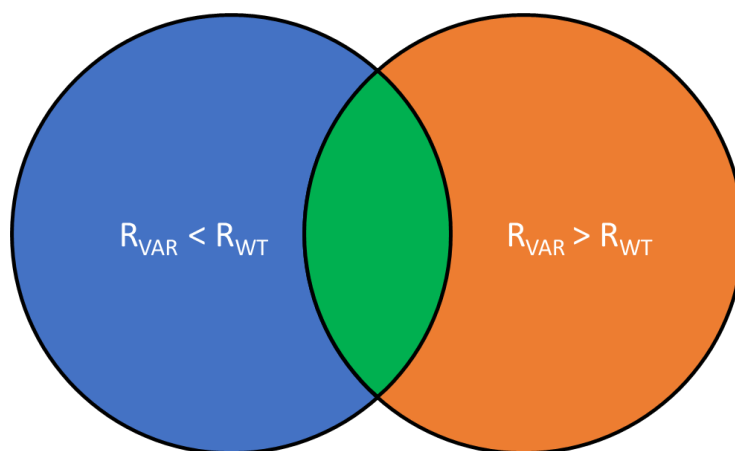


Figure 4.2: Venn diagram showing the separation of low (blue) , medium (green) and high (orange/red) MI where R_{VAR} and R_{WT} represent the residues of the variants and WT.

Machine learning model

The 12 CA and 56 CI variants along with the 99 residues composed a matrix of 68 rows and 99 columns representing the residues (independent features) along with the numerically encoded activity as a binary classifier (1 for CA and 0 for CI) for each case. Further, the dataset was split and shuffled into training and test sets using an 80/20 ratio. To handle the imbalance in the dataset, synthetic minority oversampling (SMOTE) was done, using the `smotefamily`²⁷¹ library in R²⁶⁶⁻²⁶⁸, on the training set only. We took advantage of the issue with multicollinearity by removing the independent features with high multicollinearity using variance inflation factor (VIF). The greater the magnitude of VIF found from a pair of residues has a greater potential those residue pair will likely increase multicollinearity. Such that, if a feature, R1 is correlated to features R3, R5, R8, we only keep the one that has the larger to the target (dependent feature which is an experimental activity) and remove the rest of independent features in the dataset.

The training data was input into a leave-one-out cross-validation (LOOCV), using `cv.glmnet` function from the `glmnet`²⁷⁰ library in R²⁶⁶⁻²⁶⁸, with the number of folds being equal to the number of cases from the training data set. To evaluate the LOOCV the misclassification error was determined for the datum that was left out.

$$CV(\lambda) = \frac{1}{N} \sum_{i=1}^N L(y_i, \lambda^{(-K(i))}(x_i)) \quad \text{eq. 4.4}$$

Equation 4.4 shown above represents the leave-one-out cross-validation^{272, 295-299}. A LOOCV fits a model by repeating N times for each datum in the training set. Since having imbalanced data and using SMOTE^{271, 300} on the training data this method aims to reduce the misclassification rate as well as bias and randomness from the binary classification. The maximum variance for the prediction can be found by using K=N, and therefore, represents the extrema of the data. Both a lasso (L1) and ridge (L2) regression^{272, 295-299} were evaluated, with similar results. However, the L1 regression was used for the model as it imposes a size constraint on the tuning parameter λ and reduces multicollinearity. The model was then used to evaluate the test as well as the training set for accuracy and misclassification using a confusion matrix along with the F1-score, recall, and precision metrics.

Results and Discussion

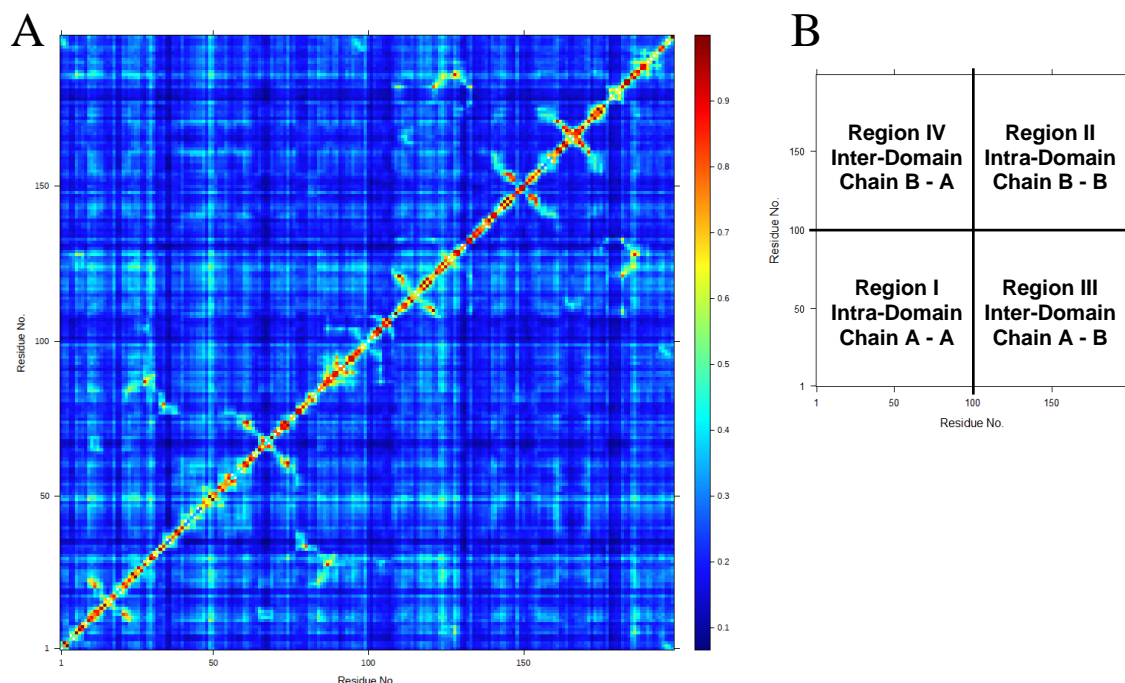


Figure 4.3: (A) Representation of the pairwise MI matrix of $C\alpha$ atoms in WT of HIV-1 protease bound to MA/CA substrate. The x and y axis are the residue numbers, and the points represent the normalized MI value between 0 and 1. (B) Shows the regions of intra-domain and inter-domain correlations found in Figure 4.2A.

The pairwise MI matrix of the WT structure is shown in Figure 4.3A, where we can see a grid like pattern of information and junctions as transmissions are passed from one residue to another. Each point is colored differently based on the magnitude of MI and it appears that most of the pairwise MI values fall in the range of 0.1 to 0.5. Residue pairs found to have highest MI are found along the diagonal but there is evidence of moderate and high MI observed from distal residues throughout inter-domain region, Figure 4.2B. For example, residues 2 and 3 interact with residue 49 which are found in distinctly

different positions of the protein. Residues 2 and 3 are found in the hinge region and on the other hand, residue 49 is found in the flap which is a predominant residue ensuring the flap is closed.

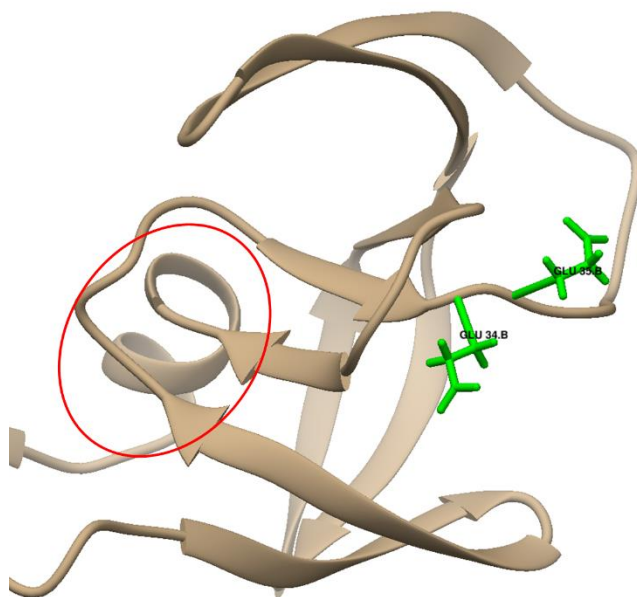


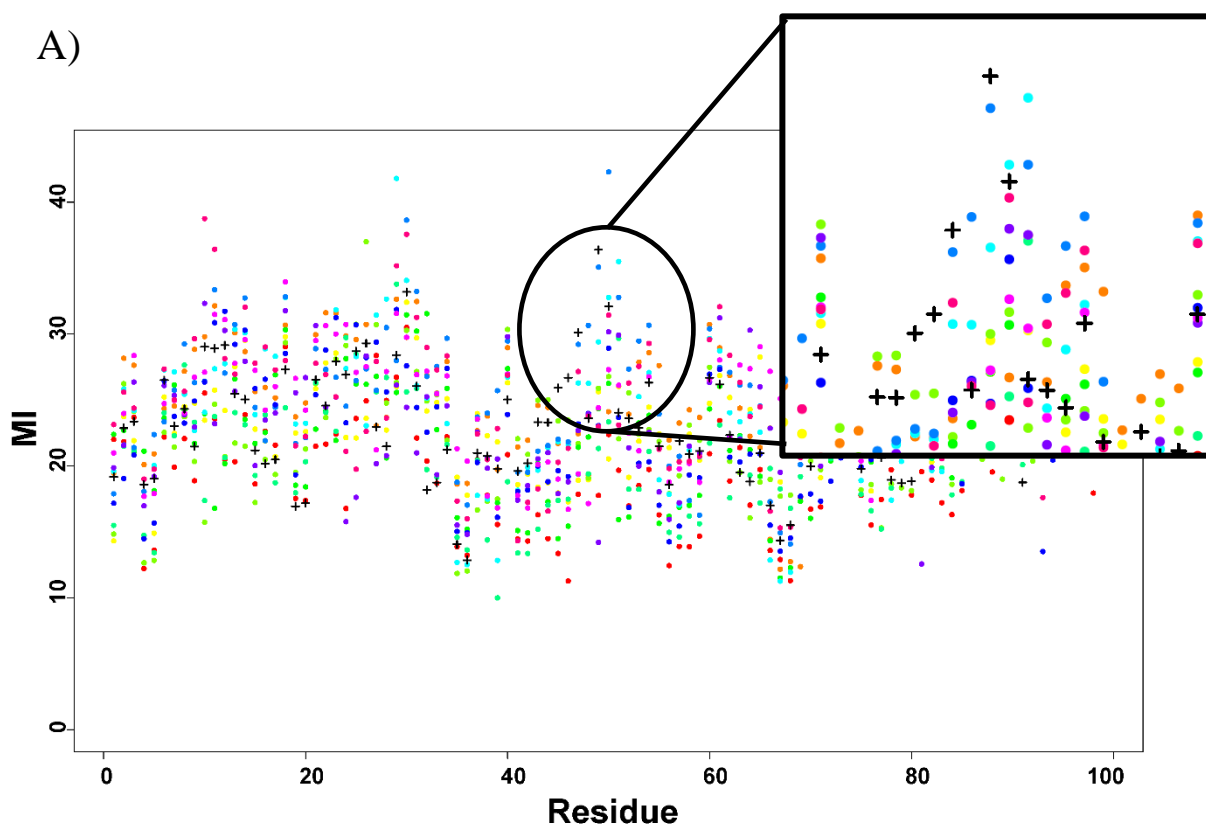
Figure 4.4: Representation of a single chain of HIV-1 protease highlighting the position of residues 34 and 35 with respect to the alpha-helix

Focusing on the inter-domain region, Figure 4.2A, there are a few spots with high MI that are found within residues of the alpha helix and non-neighboring residues. For example, residues 34 and 35, which most of time, show very low MI with other residues; however, they make a stronger interaction or higher MI with the residues of the alpha helix. Residues 34 and 35 are part of a coil section in the elbow region that passes between two neighboring beta sheets, of which, one is connected to the alpha helix, Figure 4.3. This communication/interaction between those distant residues might indicate that these residues may be functional residues on the allosteric pathway of the protein structure.

The total MI of each residue across the MD production runs was calculated for the HIV-1 protease WT and each variant structure (CA and CI variants described in Figure 4.4 A&B, respectively). In figure 4.4A, there are 12 distinct colors representing the 12 CA variants structures. In Figure 4.4B, 56 diverse colors are used to describe the 56 CI variants structures. In both figures, the total MI of each residue of the WT structure is identified by the black cross to distinguish the changes from that of the variant's structures. Comparing the total MI of each residue in the WT versus VT (variant type) shows that the variants (CA/CI) cause fluctuations/changes in the information that each residue carries throughout the MD simulation. Even though there is only a single point substitution in a VT structure, the changes found span across every residue (there are a few residues that do not show as much change compared to WT that is discussed further in the following section). These difference in the VT structures MI further supports the allosteric effect of non-neighboring residues. The allosteric communication in a VT structure can lead to the disruption of the evolutionary dynamics observed in the WT and as a result alter the catalytic activity/inactivity of the protein structure.

It is observable that the residues from certain residues of a VT structure can also carry less or more information than that of the same residue in WT structure. A closer look at these figures, in Figure 4.5A we can see that there are some residues such as residues 47-50 that have less MI than the same residues for the WT almost for all the 12 cases of CA variants. On the other hand, in Figure 4.5A, there are some residues such as residues 32-34 that more often shown as having more MI than that of the same residues for the WT

for essential all of the 12 CA variants. This behavior is seen with regard to the CI variants as well, Figure 4.5B. For instance, residues 47-50 show less MI in majority of CI variant structures compared to the same residues for the WT. The opposite is seen with residues 78-80 where almost all of the 56 CI variant structures show more MI than that seen in the same residues for the WT. These residues might be on the allosteric pathway and the information sent and received by them might have the greater influence on the protein function with regard to a single point substitution. Further analysis on these observations is investigated in greater detail in the follow sections.



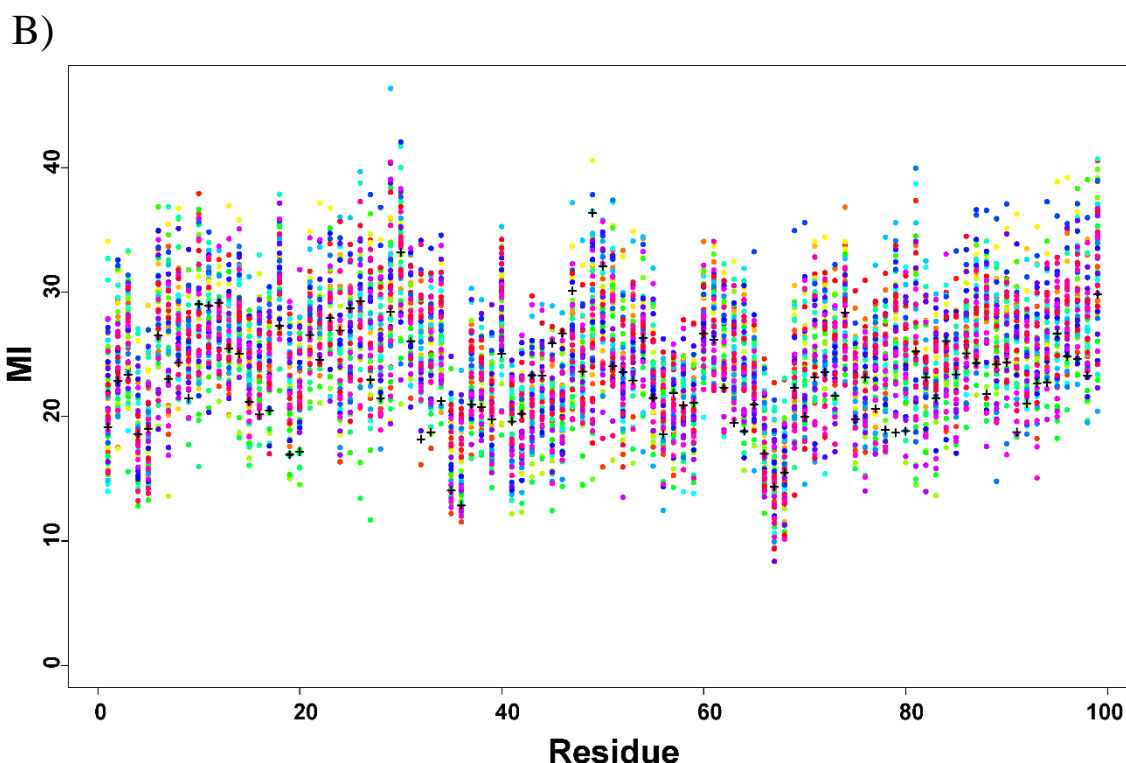
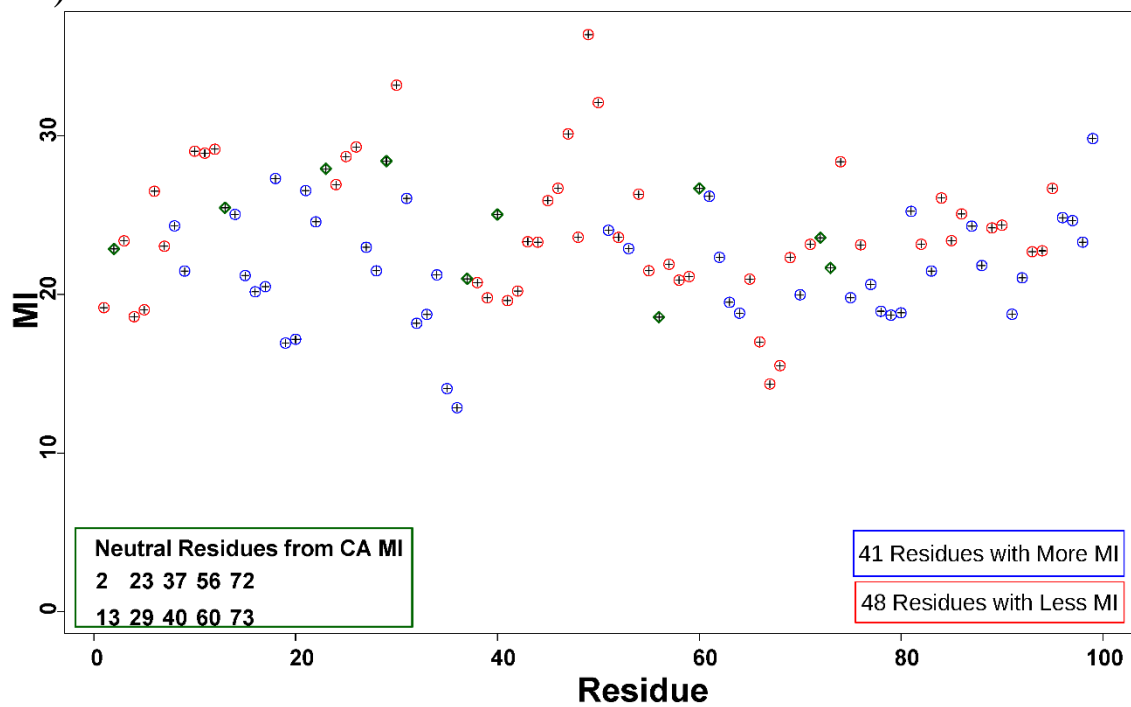


Figure 4.5: Total MI of each residue for the WT and **A)** the active variants as well as **B)** the inactive variants. The WT is represented as a black cross and each variant is represented by a unique color, 12 colors for the active and 56 colors for the inactive.

The total MI from the WT of each residue is depicted in Figure 4.6 A&B in which the residues are categorized into 3 classes based on the observations from Figures 4.5 A&B. For any residue, the total MI in variant structures (Figure 4.5 A&B) compared to the one in WT and if the residue has less MI than the WT for at least 55% of the total variant cases, it can be categorized as low MI. Residues with high MI are found if at least 55% of the variant cases for a residues are shown to have more MI than the WT. Whereas a residue was found to be neutral if the total MI for the variants cases of a residue exceeds that of the WT for less than 55% and greater than 45% of the variants. The focus of Figure 4.6 A&B is to identify residues that can act as potential allosteric hotspots due to their increase

or decrease in MI based on the impact from a single point substitution. Figure 4.6 A&B shows the total MI of each residue of the WT with the residues found in the corresponding variant cases that are more often greater than (Blue), less than (Red) or neutral (Green) with respect to the MI from the WT. The residue numbers are represented on the x-axis and the total MI is shown along the y-axis. Figure 4.6A focuses exclusively on the CA variant cases and shows that 48 residues from the CA variants are more often less, 41 of the residues are found to have more and 10 are neutral MI with respect to that of the WT. The focus of Figure 4.6B is on the CI variant cases and shows that only 28 of the residues from the CI variants are more often less, 55 of the residues are found to have more and 16 are neutral MI compared to that of the WT. The neutral residues vary slightly between the CA and CI variants, though residues 2, 13 and 60 are found to be neutral in either case. Each of the common residues are found in different locations of the structure with residue 2 being part of the hinge, residue 13 is located within the cantilever while residue 60 is located within the elbow. Overall, we see that the CI variants more often cause an increase in information with respect to the MI from the WT. This suggests that the communication sent or received by the substituted residues or residues they interact with most are highly important for the function of the structure.

A)



B)

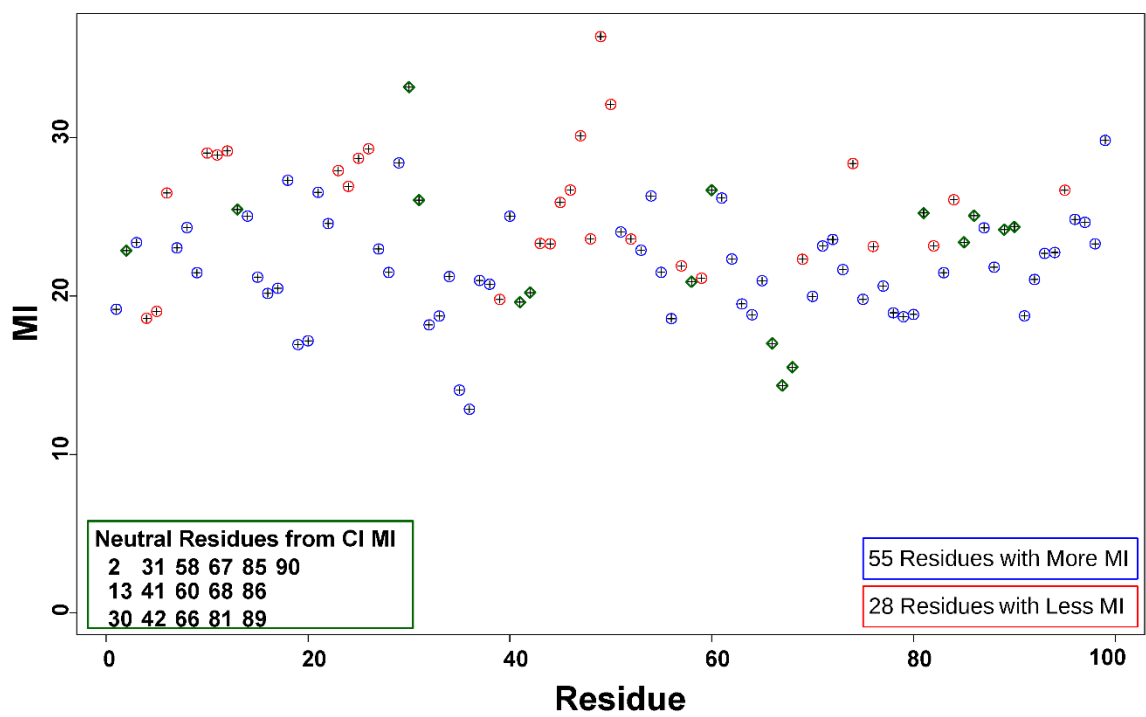


Figure 4.6: Comparison of the total MI for the WT and **A)** the active variants as well as **B)** the inactive variants. Residues where the total MI of a variant's residue is found to be more often greater than that of the WT are shown as red. Residues where the total MI of a variant's residue is found to be more often less than that of the WT are shown as blue. Residues where the total MI of a variant's residue is found to be more often similar to that of the WT are shown as green.

Here, we take advantage of a machine learning (ML) technique to investigate whether one can predict if a single point substitution causes activity/inactivity of the protein. The dataset is composed of 68 data points with each row belonging to one of the 12 CA and 56 CI variants. There are 99 columns (independent features) representing the residues in a single chain of HIV-1 protease and the value of each of these columns is the total MI rate of the corresponding residue with the other residues (except with itself). Also, we have a binary classification problem with the target value of 1 for the CA variants and 0 for the CI variants. The dataset (shuffled) was split (randomly) into training and testing sets with 80 and 20 percent of datapoints in each, respectively. To remove the issue with potential overfitting from having a large number of features in our somewhat small database (58 CI and 12 CA variants) we take advantage of the multicollinearity concern where there is a significant dependency between these independent features. Removing the multicollinearity issue from the training dataset leads us to a smaller set of 14 independent features, Figure S-4.2H. In order to further reduce the potential bias in the system due to imbalanced dataset (58 total variants), synthetic minority oversampling (SMOTE) was implemented on the training data only to increase the number of active cases from 12 to 42. Using SMOTE allows for a less bias distribution of different classes in the dataset.

The training data was fit to the logistic regression using a leave one out cross validation (LOOCV) technique. The LOOCV is used to complement the SMOTE and feature reduction techniques to adequately reduce bias and can be a good choice due to our small dataset. The LOOCV was applied using a binomial algorithm with a lasso penalty, the same calculation was done with a ridge penalty that provided comparable results. A LOOCV can be used to find a minimum lambda that can be applied to the prediction algorithm as the normalization factor for performing the logistic regression calculation on the unclassified data. The misclassification error with the $\log(\text{Lambda})$ on the x-axis and misclassification error on the y-axis, the vertical line representing the best lambda from the cross validation (Figure S-4.3). The fitted model was then validated on the training set and used to evaluate the performance of the model on the unseen datapoints from the test set.

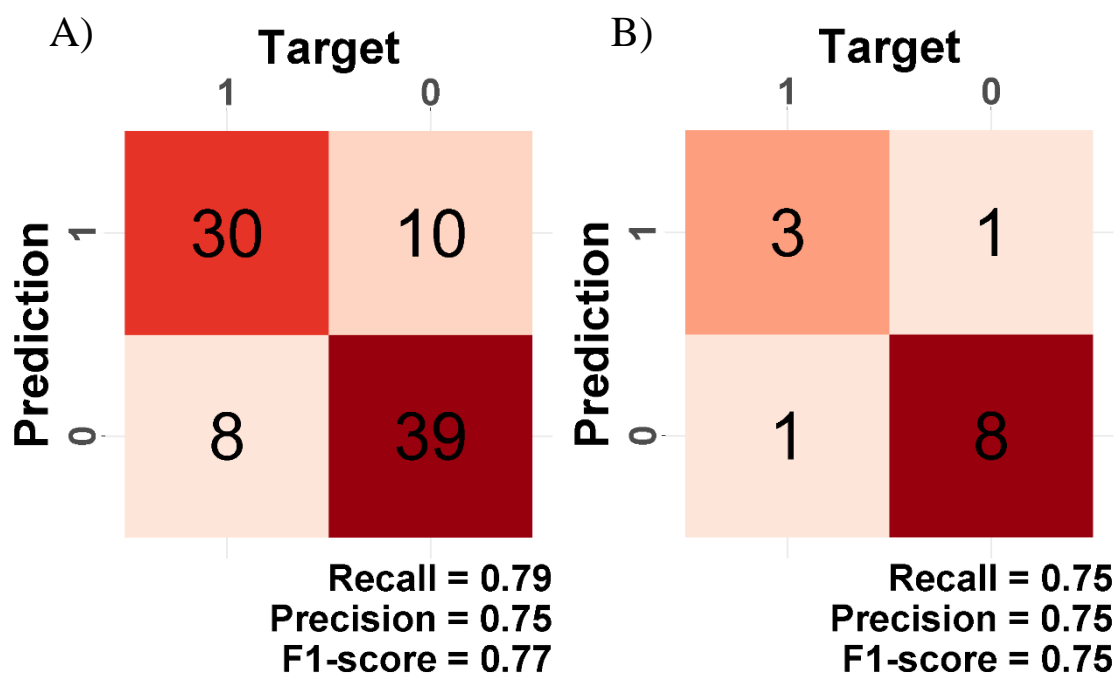


Figure 4.7: Confusion matrix of the Logistic Regression model performance **A)** the performance result of training set **B)** the performance result of test set with the recall, precision and F1-score shown at the bottom right.

The performance of the trained model for the LOOCV logistic regression is shown in Figure 4.7A and the test set Figure 4.7B along with the corresponding metrics such as F1-score, recall and precision shown at the bottom right of the figure. The result shows that the model successfully can predict 8 CI variants out of 9 CI variants and 3 CA variants out of 4 CA variants. The misclassification belongs to 1 false negative (FN) and 1 false positive (FP). However, the overall performance of the model offers a good prediction of both classes specifically the minority (active) cases.

To improve the performance of the result, the XGBoost which in an ensemble model was implemented on the same train dataset and the result is shown in Figure S-4.4A and the test set Figure S-4.4B along with the corresponding metrics such as F1-score, recall and precision shown at the bottom right of the figure. However, the performance of the XGBoost for this use case is not good, as it compared to the one from the logistic regression. Overall, we developed a logistic regression algorithm that enables us to predict successfully whether a single point substitution can be counted as a CI or CA variant.

Table 4.1: 8 runs of the logistic regression model using different random seeds.

	Train F1-score	Test F1-score	Train Recall	Test Recall
Set 1	0.709	0.500	0.718	0.500
Set 2	0.747	0.571	0.739	0.500
Set 3	0.790	0.571	0.780	0.667
Set 4	0.759	0.571	0.769	0.667
Set 5	0.774	0.571	0.750	0.500
Set 6	0.690	0.571	0.714	0.500
Set 7	0.792	0.667	0.745	0.667
Set 8	0.769	0.750	0.789	0.750

Conclusion

Allosteric signaling is a complex phenomenon that occurs within a protein structure that is uniquely shaped by the features of the amino acids that make up the structure. Relating the unique nature of the amino acids responsible for the change in activity requires in-depth examination of how the information is transformed between residues of the variant

cases. The total MI of the residue positions responsible for the change in activity offer supporting details to identify the relationship between dynamics and function of the variant. It can be observed that as the MI for a residue increases the amount of information disrupts the signals sent or received from the residues responsible for the function of the structure and results in an CI variant. Interestingly, if the information is suppressed more, as observed with the CA variants, for specific residues the function of the enzyme is moderately improved, given the CA variants observed in this study all have equivalent or greater activity than the WT. The residues seen to receive more or less information than what is commonly observed from the WT may be associated with allosteric “hotspots” and/or disrupt the allosteric pathway/s within a protein. These allosteric pathways and/or the allosteric “hotspots” can be exploited as a target during drug trials as they can adversely impact the function of the protein. Essentially, the communication from the residues responsible for the function of the enzyme can be tuned based on the modifications that occur depending on the allosteric behavior of the residue site.

The approaches taken here allowed us to describe the dynamics of HIV-1 protease by evaluating the non-linear dependency identified from the interactions of each pair of residue positions, using Mutual Information (MI). In this study we look to uncover if the communication sent or received by a substituted residue or the residue’s they interact with most can act as a constraint on the function of HIV-1 protease. The machine learning (ML) algorithms offer a substantial assessment connecting the non-linear pairwise interactions of HIV-1 protease variants with their experimentally determined activity. Furthermore,

using the ML algorithms to predict the impact caused by a single point substitution offers a suitable method for solving binary-classification problem of identifying if a variant will be active or inactive from the MI obtained from the MD simulations of the HIV-1 protease variants. In general, the non-linear correlations associated with the dynamics we've examined here show there is a connection to the activity of HIV-1 protease and can aid scientist to better understanding of the allosteric mechanisms in other proteins.

APPENDICES

Appendix A

Supplementary figures for Chapter II

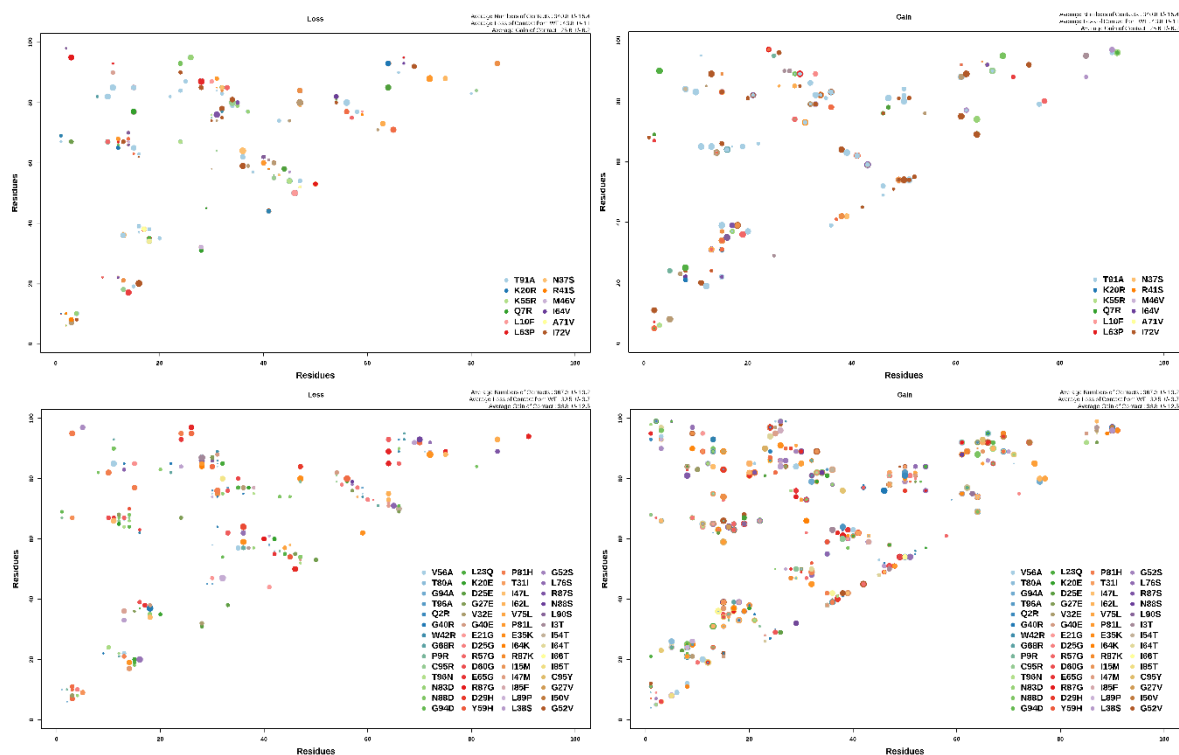


Figure S-2.1: Increased frequency of contacts.

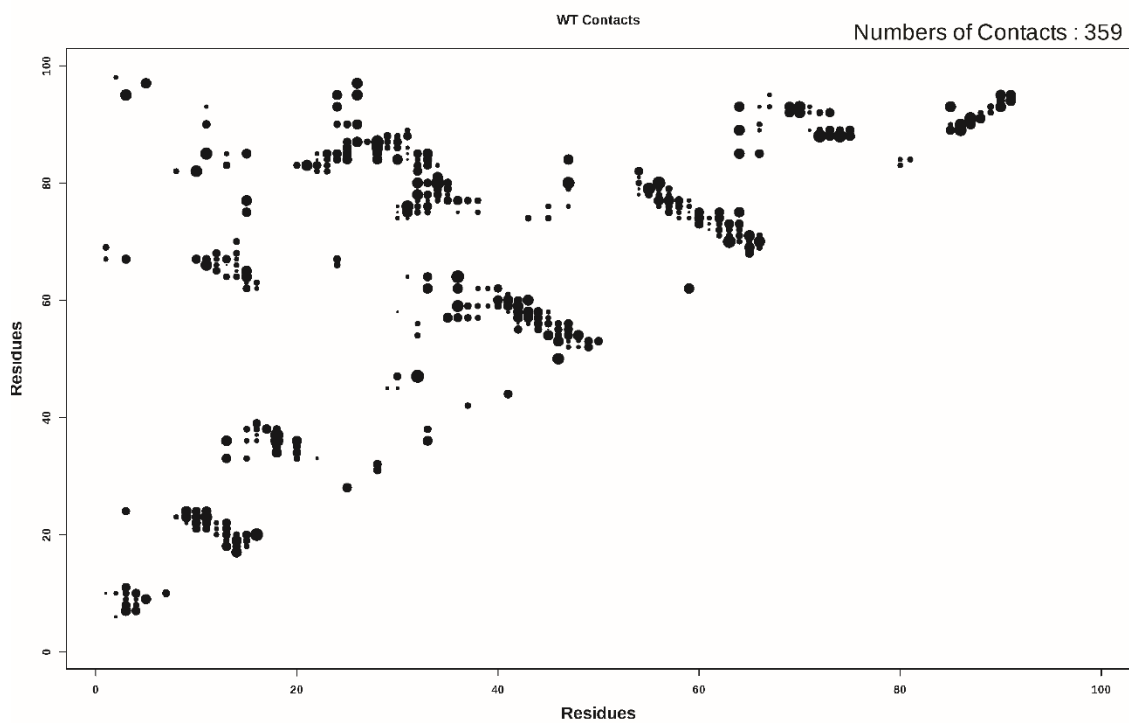


Figure S-2.2: Gain of contacts reference distance for the residues of the WT.

Table S-2.1: Change in binding free energy upon substitution (ddG) using MMPBSA and Prodigy^{105, 110, 227-235}.

Binding Free Energy ddG															
Variant	dG ¹	dG ²	dG ³	<dG ^{1,2} >	<dG ^{1,3} >	<dG ^{2,3} >	<dG ^{1,2,3} >	ddG ¹	ddG ²	ddG ³	<ddG ^{1,2} >	<ddG ^{1,3} >	<ddG ^{2,3} >	<ddG ^{1,2,3} >	Activity
WT	-19.1	-14.38	-24.72	-16.74	-21.91	-19.55	-19.4	0	0	0	0	0	0	0	1
I64V	-18.9	8.23	-24.27	-5.33	-21.59	-8.02	-11.65	0.2	22.61	0.45	11.41	0.33	11.53	7.75	1
I72V	-18.9	-16.81	-22.8	-17.86	-20.85	-19.81	-19.5	0.2	-2.43	1.92	-1.11	1.06	-0.25	-0.1	1
K20R	-18.4	-18.15	-24.1	-18.27	-21.25	-21.13	-20.22	0.7	-3.77	0.62	-1.53	0.66	-1.57	-0.82	1
A71V	-17.3	-1.52	-23.01	-9.41	-20.15	-12.26	-13.94	1.8	12.86	1.72	7.33	1.76	7.29	5.46	1
L63P	-17.2	4.78	-26.28	-6.21	-21.74	-10.75	-12.9	1.9	19.16	-1.56	10.53	0.17	8.8	6.5	1
K55R	-16.6	0.86	-22.8	-7.87	-19.7	-10.97	-12.84	2.5	15.25	1.93	8.87	2.21	8.59	6.56	1
T91A	-16.5	-11.99	-21.97	-14.24	-19.24	-16.98	-16.82	2.6	2.39	2.75	2.5	2.68	2.57	2.58	1
Q7R	-16.3	1.68	-25.94	-7.31	-21.12	-12.13	-13.52	2.8	16.06	-1.22	9.43	0.79	7.42	5.88	1
N37S	-16.3	-7.2	-23.8	-11.75	-20.05	-15.5	-15.76	2.8	7.18	0.93	4.99	1.86	4.06	3.64	1
R41S	-15.9	-13.86	-20.56	-14.88	-18.23	-17.21	-16.77	3.2	0.52	4.17	1.86	3.68	2.35	2.63	1
L10F	-15.8	-22.64	-22.55	-19.22	-19.18	-22.6	-20.33	3.3	-8.26	2.17	-2.48	2.74	-3.04	-0.93	1
M46V	-15.6	-15.05	-21.77	-15.32	-18.68	-18.41	-17.47	3.5	-0.67	2.96	1.42	3.23	1.14	1.93	1
D25G	-19.7	-17.37	-24.58	-18.53	-22.14	-20.97	-20.55	-0.6	-2.98	0.14	-1.79	-0.23	-1.42	-1.15	0
Y59H	-19.5	-17.88	-24.9	-18.69	-22.2	-21.39	-20.76	-0.4	-3.5	-0.18	-1.95	-0.29	-1.84	-1.36	0
L23Q	-19.2	-9.48	-24.44	-14.34	-21.82	-16.96	-17.71	-0.1	4.9	0.28	2.4	0.09	2.59	1.7	0
I3T	-18.7	-6.96	-24.12	-12.83	-21.41	-15.54	-16.59	0.4	7.42	0.6	3.91	0.5	4.01	2.81	0
G94A	-18.6	-15.04	-25.95	-16.82	-22.27	-20.5	-19.86	0.5	-0.66	-1.23	-0.08	-0.36	-0.94	-0.46	0
D29H	-18.5	-16.02	-26.04	-17.26	-22.27	-21.03	-20.19	0.6	-1.63	-1.32	-0.52	-0.36	-1.48	-0.78	0
I64T	-18.4	-20.19	-23.76	-19.29	-21.08	-21.97	-20.78	0.7	-5.81	0.96	-2.55	0.83	-2.42	-1.38	0
K20E	-18.2	-7.48	-24.96	-12.84	-21.58	-16.22	-16.88	0.9	6.9	-0.23	3.9	0.33	3.33	2.52	0
R87S	-18.2	-7.35	-25.91	-12.77	-22.05	-16.63	-17.15	0.9	7.03	-1.18	3.97	-0.14	2.92	2.25	0
I54T	-18.2	-11.86	-27.78	-15.03	-22.99	-19.82	-19.28	0.9	2.52	-3.05	1.71	-1.08	-0.27	0.12	0
G94D	-18.1	-6.31	-21.78	-12.2	-19.94	-14.04	-15.4	1	8.07	2.95	4.54	1.97	5.51	4.01	0
V56A	-17.9	-17.54	-26.01	-17.72	-21.96	-21.78	-20.48	1.2	-3.16	-1.29	-0.98	-0.04	-2.22	-1.08	0
W42R	-17.9	-22.22	-21.98	-20.06	-19.94	-22.1	-20.7	1.2	-7.84	2.75	-3.32	1.97	-2.54	-1.3	0
N83D	-17.9	-6.51	-23.02	-12.2	-20.46	-14.76	-15.81	1.2	7.88	1.71	4.54	1.45	4.79	3.59	0
T96A	-17.8	-17.38	-23.86	-17.59	-20.83	-20.62	-19.68	1.3	-3	0.86	-0.85	1.08	-1.07	-0.28	0
D60G	-17.7	-7.76	-20.37	-12.73	-19.03	-14.07	-15.28	1.4	6.62	4.36	4.01	2.88	5.49	4.12	0
L89P	-17.6	-22.63	-21.51	-20.12	-19.55	-22.07	-20.58	1.5	-8.25	3.22	-3.38	2.36	-2.52	-1.18	0
R57G	-17.5	-18.54	-24.98	-18.02	-21.24	-21.76	-20.34	1.6	-4.15	-0.25	-1.28	0.67	-2.2	-0.94	0
P9R	-17.3	-8.31	-26.23	-12.8	-21.76	-17.27	-17.28	1.8	6.07	-1.5	3.94	0.15	2.29	2.12	0
C95R	-17.3	34.42	-29.53	8.56	-23.41	2.44	-4.14	1.8	48.8	-4.8	25.3	-1.5	22	15.27	0

P81L	-17.3	-8.77	-26.63	-13.03	-21.97	-17.7	-17.57	1.8	5.61	-1.91	3.71	-0.05	1.85	1.83	0
G68R	-17.2	-8.72	-24.19	-12.96	-20.69	-16.45	-16.7	1.9	5.66	0.54	3.78	1.22	3.1	2.7	0
R87G	-17.2	-13.78	-24.79	-15.49	-20.99	-19.28	-18.59	1.9	0.61	-0.06	1.25	0.92	0.27	0.81	0
P81H	-17.2	-7.8	-26.13	-12.5	-21.66	-16.96	-17.04	1.9	6.59	-1.4	4.24	0.25	2.59	2.36	0
G27V	-17.2	-17.39	-26.75	-17.29	-21.97	-22.07	-20.45	1.9	-3.01	-2.02	-0.55	-0.06	-2.51	-1.04	0
N88D	-17.1	-10.85	-24.5	-13.97	-20.8	-17.67	-17.48	2	3.54	0.23	2.77	1.11	1.88	1.92	0
V75L	-17	-15.61	-26.72	-16.3	-21.86	-21.16	-19.78	2.1	-1.23	-2	0.44	0.05	-1.61	-0.37	0
T31I	-16.9	-8.59	-22.95	-12.75	-19.92	-15.77	-16.15	2.2	5.79	1.78	3.99	1.99	3.78	3.25	0
E35K	-16.7	-7.01	-24.19	-11.85	-20.44	-15.6	-15.96	2.4	7.38	0.54	4.89	1.47	3.96	3.44	0
L76S	-16.7	-8.65	-23.15	-12.67	-19.93	-15.9	-16.17	2.4	5.74	1.57	4.07	1.99	3.65	3.24	0
I50V	-16.7	0.78	-22.77	-7.96	-19.74	-11	-12.9	2.4	15.16	1.95	8.78	2.18	8.56	6.5	0
I85T	-16.3	10.38	-25.16	-2.96	-20.73	-7.39	-10.36	2.8	24.76	-0.44	13.78	1.18	12.16	9.04	0
E21G	-16.2	-14.21	-22.98	-15.2	-19.59	-18.59	-17.8	2.9	0.17	1.74	1.54	2.32	0.96	1.61	0
I66T	-16.2	-0.38	-25.76	-8.29	-20.98	-13.07	-14.11	2.9	14	-1.04	8.45	0.93	6.48	5.29	0
E65G	-16.1	2.54	-27.45	-6.78	-21.78	-12.46	-13.67	3	16.92	-2.73	9.96	0.14	7.1	5.73	0
I85F	-16.1	-12.16	-23.08	-14.13	-19.59	-17.62	-17.12	3	2.22	1.64	2.61	2.32	1.93	2.29	0
G40R	-16	-29.76	-27.66	-22.88	-21.83	-28.71	-24.47	3.1	-15.37	-2.94	-6.14	0.08	-9.16	-5.07	0
D25E	-16	-18.14	-23.86	-17.07	-19.93	-21	-19.34	3.1	-3.76	0.86	-0.33	1.98	-1.45	0.07	0
G52V	-15.9	-11.81	-21.75	-13.85	-18.82	-16.78	-16.48	3.2	2.57	2.98	2.89	3.09	2.78	2.92	0
T80A	-15.7	11.66	-25.5	-2.02	-20.6	-6.92	-9.85	3.4	26.04	-0.78	14.72	1.31	12.63	9.55	0
G27E	-15.7	-26.34	-20.54	-21.02	-18.12	-23.44	-20.86	3.4	-11.96	4.18	-4.28	3.79	-3.89	-1.46	0
I64K	-15.7	-3.19	-23.39	-9.45	-19.55	-13.29	-14.1	3.4	11.19	1.33	7.29	2.37	6.26	5.31	0
I47M	-15.7	-16.82	-25.07	-16.26	-20.39	-20.95	-19.2	3.4	-2.44	-0.35	0.48	1.53	-1.39	0.2	0
C95Y	-15.7	9.06	-23.4	-3.32	-19.55	-7.17	-10.01	3.4	23.44	1.33	13.42	2.36	12.38	9.39	0
I47L	-15.5	12.64	-22.88	-1.43	-19.19	-5.12	-8.58	3.6	27.03	1.84	15.31	2.72	14.43	10.82	0
L38S	-15.4	22.87	-24.58	3.74	-19.99	-0.85	-5.7	3.7	37.26	0.15	20.48	1.92	18.7	13.7	0
T96N	-15.1	41.56	-27.65	13.23	-21.38	6.95	-0.4	4	55.94	-2.93	29.97	0.54	26.51	19.01	0
L90S	-15.1	-11.23	-22.74	-13.16	-18.92	-16.98	-16.36	4	3.15	1.99	3.58	2.99	2.57	3.05	0
G40E	-14.9	-15.43	-23.99	-15.17	-19.45	-19.71	-18.11	4.2	-1.05	0.73	1.58	2.47	-0.16	1.29	0
I62L	-14.8	-16.01	-23.91	-15.4	-19.36	-19.96	-18.24	4.3	-1.63	0.81	1.34	2.56	-0.41	1.16	0
I15M	-14.7	-9.32	-29.91	-12.01	-22.3	-19.61	-17.98	4.4	5.06	-5.18	4.73	-0.39	-0.06	1.43	0
R87K	-14.5	-22.97	-23.8	-18.73	-19.15	-23.39	-20.42	4.6	-8.58	0.92	-1.99	2.76	-3.83	-1.02	0
Q2R	-13.8	0.14	-24.87	-6.83	-19.33	-12.36	-12.84	5.3	14.52	-0.14	9.91	2.58	7.19	6.56	0
G52S	-13.7	-0.3	-28.1	-7	-20.9	-14.2	-14.03	5.4	14.08	-3.37	9.74	1.01	5.36	5.37	0
N88S	-13.5	-2.51	-21.23	-8	-17.37	-11.87	-12.41	5.6	11.88	3.49	8.74	4.55	7.68	6.99	0
V32E	-13	20.88	-28.51	3.94	-20.76	-3.82	-6.88	6.1	35.26	-3.79	20.68	1.16	15.74	12.52	0

Table S-2.2: Change in protein-protein binding free energy upon substitution (ddG) using Prodigy^{105, 110, 227-235}.

Protein-Protein ddG							
Variant	dG	ddG	Activity				
WT	-11.6	0	1	I85T	-11.6	0	0
L10F	-12.4	0.8	1	C95Y	-11.6	0	0
K55R	-12.1	0.5	1	D29H	-11.5	-0.1	0
Q7R	-12.1	0.5	1	L89P	-11.5	-0.1	0
L63P	-12.1	0.5	1	I3T	-11.5	-0.1	0
N37S	-12.1	0.5	1	G52V	-11.5	-0.1	0
R41S	-12.1	0.5	1	V56A	-11.4	-0.2	0
I72V	-11.9	0.3	1	T96N	-11.4	-0.2	0
K20R	-11.6	0	1	N88D	-11.3	-0.3	0
I64V	-11.3	-0.3	1	L76S	-11.3	-0.3	0
T91A	-11.2	-0.4	1	G94A	-11.2	-0.4	0
A71V	-11.1	-0.5	1	T96A	-11.2	-0.4	0
M46V	-10.7	-0.9	1	W42R	-11.2	-0.4	0
D25G	-13.3	1.7	0	Y59H	-11.2	-0.4	0
G40R	-12.6	1	0	I64T	-11.2	-0.4	0
N83D	-12.6	1	0	C95R	-11.1	-0.5	0
D60G	-12.6	1	0	V32E	-11.1	-0.5	0
R87G	-12.2	0.6	0	E21G	-11.1	-0.5	0
I54T	-12.2	0.6	0	R57G	-11.1	-0.5	0
K20E	-12.1	0.5	0	N88S	-11.1	-0.5	0
L23Q	-12	0.4	0	D25E	-11	-0.6	0
T31I	-12	0.4	0	P81L	-10.8	-0.8	0
G40E	-11.9	0.3	0	L38S	-10.8	-0.8	0
P81H	-11.9	0.3	0	G27V	-10.8	-0.8	0
I62L	-11.9	0.3	0	E35K	-10.7	-0.9	0
I47M	-11.9	0.3	0	T80A	-10.6	-1	0
G52S	-11.9	0.3	0	E65G	-10.6	-1	0
P9R	-11.8	0.2	0	I15M	-10.6	-1	0
V75L	-11.8	0.2	0	I66T	-10.5	-1.1	0
I50V	-11.8	0.2	0	I85F	-10.4	-1.2	0
G94D	-11.7	0.1	0	G68R	-10.2	-1.4	0
G27E	-11.7	0.1	0	I64K	-10.1	-1.5	0
L90S	-11.7	0.1	0	Q2R	-9.6	-2	0
I47L	-11.6	0	0	R87K	-8.7	-2.9	0
R87S	-11.6	0	0				

Model	Active	Inactive
ddG	0.33	0.57

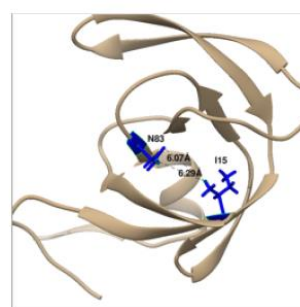
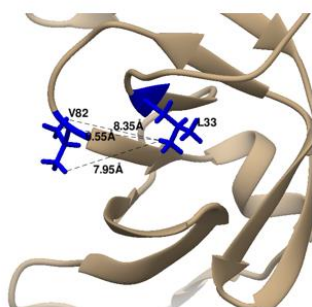
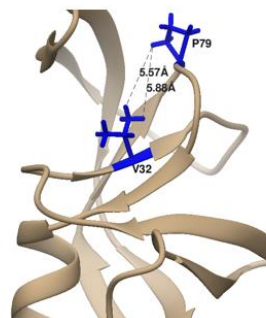
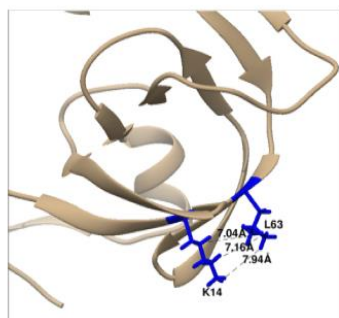


Figure S-2.3: Gain of contacts reference distance for the residues of the WT.

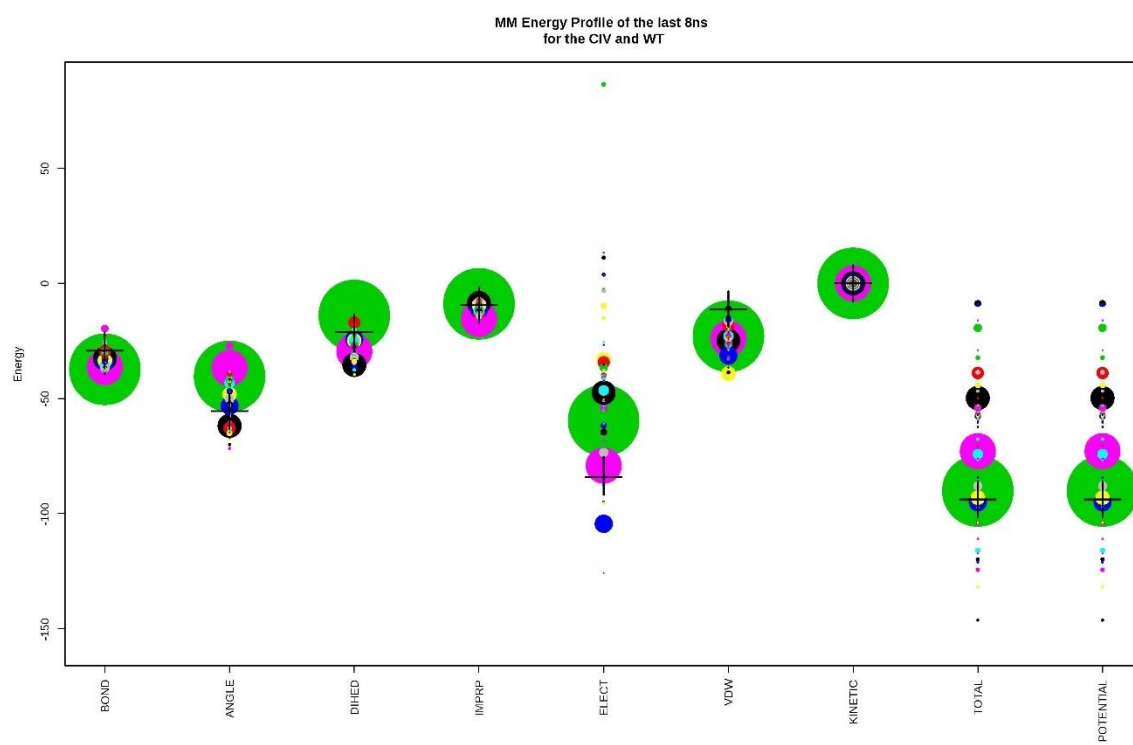
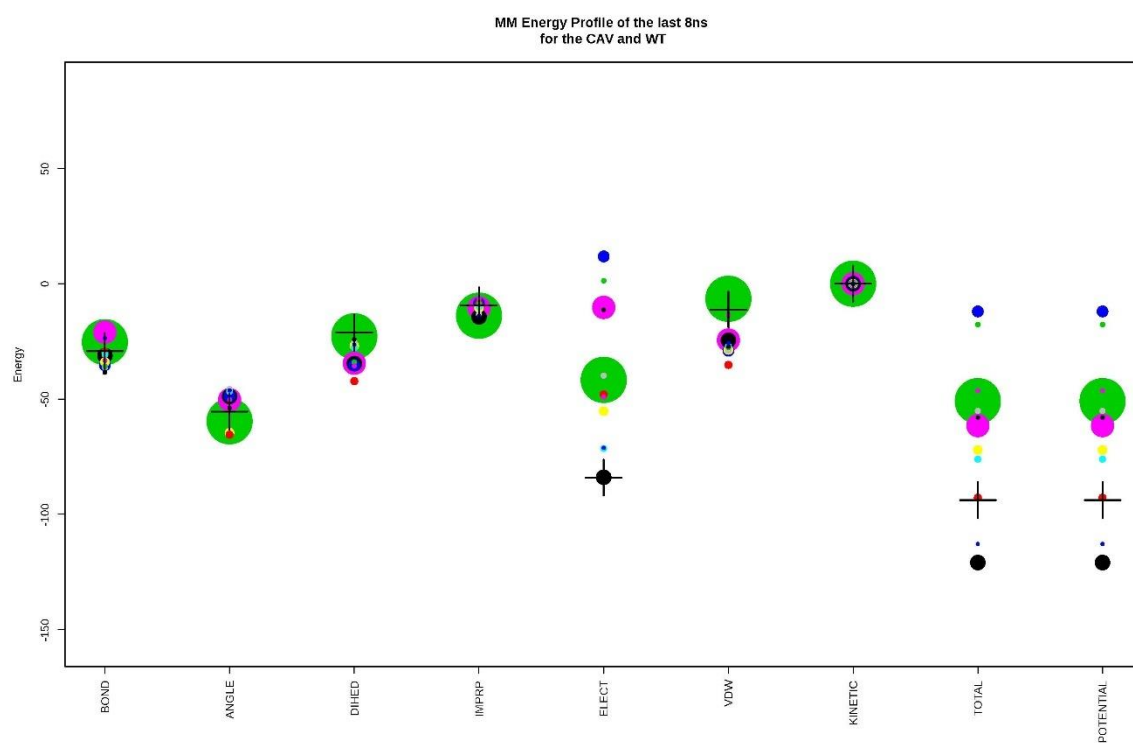


Figure S-2.4: The figures above show the MM energy (kcal/mol) along the y-axis and the individual components of the MM energy on the x-axis. Each of the CA variants (A) are represented as colored points and the WT is shown as a cross. Likewise, each of the CI variants (B) are represented as colored points and the WT is shown as a cross.

Additionally, we looked at methods to understand the impact of substitutions on the protein-protein interactions of oligomerization for HIV-1 protease using a software called Prodigy^{105, 228, 229}, developed by Vangone et.al. Prodigy predicted that about 30% of the CA variants while 57% of the CI variants have a stabilizing protein-protein interaction, Supplementary Table A. Having a stable protein-protein interface for an enzyme like HIV-1 protease is essential, however stability can be a sign that the variant is more likely to prefer a closed conformation.

Appendix B

Supplementary figures for Chapter III

Figure S-3.1:

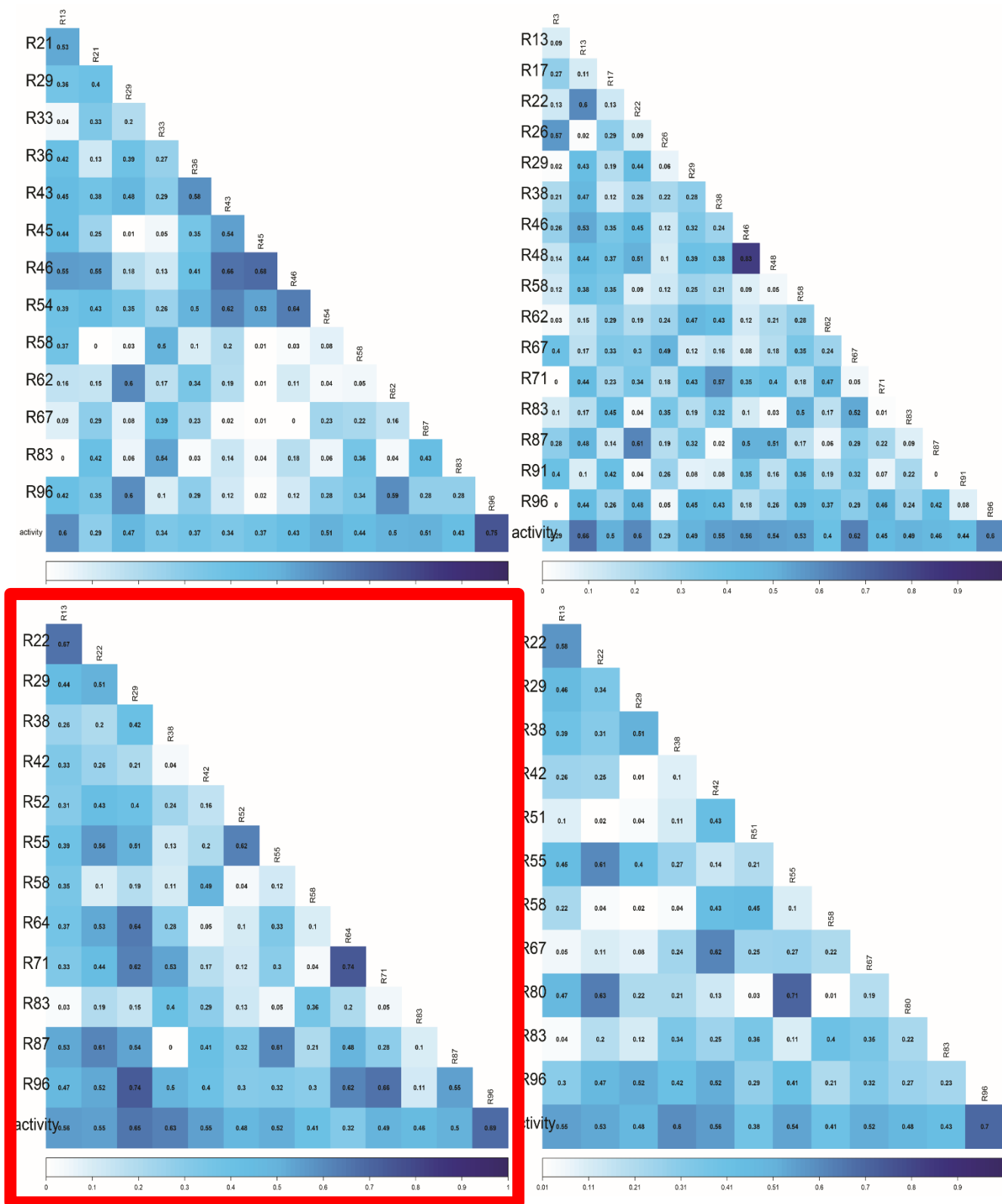
	Frequency				Total				Magnitude				
	Intra-Domain		Inter-Domain		Intra-Domain		Inter-Domain		Intra-Domain		Inter-Domain		
	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Activity
WT	2349	2502	2102	2848	618.814	-436.845	280.644	-470.905	0.263	-0.175	0.134	-0.165	1
ALA91	2242	2609	1684	3266	511.035	-315.557	186.354	-484.422	0.228	-0.121	0.111	-0.148	1
ARG20	2595	2256	1795	3155	731.297	-466.278	253.529	-552.883	0.282	-0.207	0.141	-0.175	1
ARG55	2403	2448	1764	3186	576.300	-330.426	216.957	-517.529	0.240	-0.135	0.123	-0.162	1
ARG7	2375	2476	1858	3092	518.105	-298.422	213.016	-436.614	0.218	-0.121	0.115	-0.141	1
PHE10	2995	1856	1491	3459	717.551	-219.204	180.296	-751.822	0.240	-0.118	0.121	-0.217	1
PRO63	2279	2572	1754	3196	552.106	-360.583	222.247	-495.031	0.242	-0.140	0.127	-0.155	1
SER37	2489	2362	1768	3182	550.423	-301.635	232.921	-515.877	0.221	-0.128	0.132	-0.162	1
SER41	2850	2001	1326	3624	633.096	-230.445	135.335	-656.281	0.222	-0.115	0.102	-0.181	1
VAL46	2188	2663	1821	3129	505.544	-301.614	185.046	-413.068	0.231	-0.113	0.102	-0.132	1
VAL64	2231	2620	1731	3219	535.171	-339.373	202.958	-478.940	0.240	-0.130	0.117	-0.149	1
VAL71	2294	2557	1951	2999	620.487	-399.953	237.875	-490.200	0.270	-0.156	0.122	-0.163	1
VAL72	2257	2594	1996	2954	506.088	-353.043	284.391	-493.956	0.224	-0.136	0.142	-0.167	1
ALA56	2275	2576	2050	2900	557.495	-369.709	265.578	-486.683	0.245	-0.144	0.130	-0.168	0
ALA80	2393	2458	1623	3327	600.198	-300.594	208.752	-595.527	0.251	-0.122	0.129	-0.179	0
ALA94	2089	2762	1832	3118	468.931	-320.180	192.319	-380.943	0.224	-0.116	0.105	-0.122	0
ALA96	2373	2478	1675	3275	534.937	-301.322	189.499	-486.335	0.225	-0.122	0.113	-0.148	0
ARG2	2828	2023	1490	3460	748.486	-328.017	196.344	-710.029	0.265	-0.162	0.132	-0.205	0
ARG40	2315	2536	1957	2993	695.133	-478.268	314.635	-560.209	0.300	-0.189	0.161	-0.187	0
ARG42	2399	2452	2084	2866	543.350	-335.467	264.884	-517.727	0.226	-0.137	0.127	-0.181	0
ARG68	2565	2286	1714	3236	637.466	-374.725	284.248	-694.235	0.249	-0.164	0.166	-0.215	0
ARG9	2380	2471	2103	2847	530.430	-397.894	287.763	-409.576	0.223	-0.161	0.137	-0.144	0
ARG95	2471	2380	2095	2855	641.926	-451.644	320.985	-497.837	0.260	-0.190	0.153	-0.174	0
ASN96	2547	2304	1613	3337	667.924	-342.642	217.403	-633.486	0.262	-0.149	0.135	-0.190	0
ASP83	2498	2353	1841	3109	596.975	-376.284	198.299	-431.194	0.239	-0.160	0.108	-0.139	0
ASP88	2477	2374	1737	3213	533.023	-272.447	184.535	-504.966	0.215	-0.115	0.106	-0.157	0
ASP94	2225	2626	1641	3309	484.485	-305.411	196.091	-427.150	0.218	-0.116	0.119	-0.129	0
GLN23	2521	2330	1831	3119	598.548	-359.309	225.181	-511.013	0.237	-0.154	0.123	-0.164	0
GLU20	2265	2586	2003	2947	472.906	-315.359	232.044	-385.150	0.209	-0.122	0.116	-0.131	0
GLU25	2371	2480	1697	3253	509.757	-287.608	202.792	-482.353	0.215	-0.116	0.120	-0.148	0
GLU27	2485	2366	1815	3135	579.984	-324.120	258.250	-495.316	0.233	-0.137	0.142	-0.158	0
GLU32	2725	2126	1798	3152	676.143	-301.986	208.317	-491.125	0.248	-0.142	0.116	-0.156	0
GLU40	2594	2257	1124	3826	731.923	-336.667	126.228	-760.252	0.282	-0.149	0.112	-0.199	0
GLY21	2482	2369	1751	3199	526.089	-246.493	202.752	-497.780	0.212	-0.104	0.116	-0.156	0
GLY25	2201	2650	1744	3206	479.076	-316.232	188.155	-392.120	0.218	-0.119	0.108	-0.122	0
GLY57	2192	2659	2002	2948	523.772	-377.672	296.050	-452.545	0.239	-0.142	0.148	-0.154	0
GLY60	2312	2539	1945	3005	546.108	-327.042	256.895	-476.523	0.236	-0.129	0.132	-0.159	0
GLY65	2453	2398	1970	2980	670.254	-398.678	257.337	-624.016	0.273	-0.166	0.131	-0.209	0
GLY87	2326	2525	1884	3066	589.824	-381.793	261.308	-576.937	0.254	-0.151	0.139	-0.188	0
HIS29	2454	2397	1838	3112	520.900	-334.621	218.084	-408.323	0.212	-0.140	0.119	-0.131	0
HIS59	2568	2283	1765	3185	690.029	-334.482	263.243	-702.258	0.269	-0.147	0.149	-0.220	0
HIS81	2389	2462	1873	3077	587.884	-351.049	269.792	-601.643	0.246	-0.143	0.144	-0.196	0
ILE31	2295	2556	1660	3290	545.356	-334.175	195.079	-513.633	0.238	-0.131	0.118	-0.156	0
LEU47	2636	2215	1880	3070	676.121	-376.214	309.660	-604.684	0.256	-0.170	0.165	-0.197	0
LEU62	2411	2440	1651	3299	557.718	-355.257	184.314	-486.181	0.231	-0.146	0.112	-0.147	0
LEU75	2423	2428	1695	3255	544.671	-326.709	213.619	-533.284	0.225	-0.135	0.126	-0.164	0
LEU81	2579	2272	1861	3089	590.462	-262.940	244.672	-500.191	0.229	-0.116	0.131	-0.162	0
LYS35	2241	2610	1798	3152	569.996	-406.578	234.502	-478.171	0.254	-0.156	0.130	-0.152	0
LYS64	2699	2152	1686	3264	694.399	-297.289	258.232	-736.599	0.257	-0.138	0.153	-0.226	0
LYS87	2621	2230	1564	3386	583.252	-255.884	173.075	-543.356	0.223	-0.115	0.111	-0.160	0
MET15	2426	2425	1756	3194	595.109	-319.012	249.695	-600.986	0.245	-0.132	0.142	-0.188	0
MET47	2217	2634	1975	2975	503.963	-296.641	241.023	-409.641	0.227	-0.113	0.122	-0.138	0
PHE85	2655	2196	1576	3374	796.763	-386.658	188.976	-765.816	0.300	-0.176	0.120	-0.227	0
PRO89	2486	2365	2025	2925	591.532	-362.237	312.620	-556.351	0.238	-0.153	0.154	-0.190	0
SER38	3046	1805	1510	3440	684.873	-234.970	168.171	-640.597	0.225	-0.130	0.111	-0.186	0
SER52	2320	2531	2024	2926	588.219	-413.169	307.137	-517.499	0.254	-0.163	0.152	-0.177	0
SER76	2314	2537	1769	3181	525.891	-290.875	210.214	-442.707	0.227	-0.115	0.119	-0.139	0
SER87	2590	2261	1546	3404	638.994	-370.230	211.685	-619.695	0.247	-0.164	0.137	-0.182	0
SER88	2293	2558	1768	3182	567.223	-386.749	223.582	-515.420	0.247	-0.151	0.126	-0.162	0
SER90	2440	2411	1766	3184	629.637	-435.345	262.379	-611.527	0.258	-0.181	0.149	-0.192	0
THR3	2560	2291	1348	3602	562.256	-235.998	115.606	-493.134	0.220	-0.103	0.086	-0.137	0
THR54	2948	1903	1632	3318	658.132	-235.428	198.947	-561.933	0.223	-0.124	0.122	-0.169	0
THR64	2541	2310	1825	3125	540.405	-282.651	229.576	-532.333	0.213	-0.122	0.126	-0.170	0
THR66	2552	2299	1468	3482	660.229	-338.790	174.952	-561.049	0.259	-0.147	0.119	-0.161	0
THR85	2712	2139	1609	3341	730.574	-347.446	256.767	-699.450	0.269	-0.162	0.160	-0.209	0
TYR95	2254	2597	1877	3073	514.603	-315.288	227.607	-447.043	0.228	-0.121	0.121	-0.145	0
VAL27	2279	2572	1796	3154	528.435	-324.065	200.002	-425.563	0.232	-0.126	0.111	-0.135	0
VAL50	2372	2479	2082	2868	581.034	-372.217	310.269	-540.009	0.245	-0.150	0.149	-0.188	0
VAL52	3122	1729	1578	3372	936.073	-274.003	261.750	-897.902	0.300	-0.158	0.166	-0.266	0

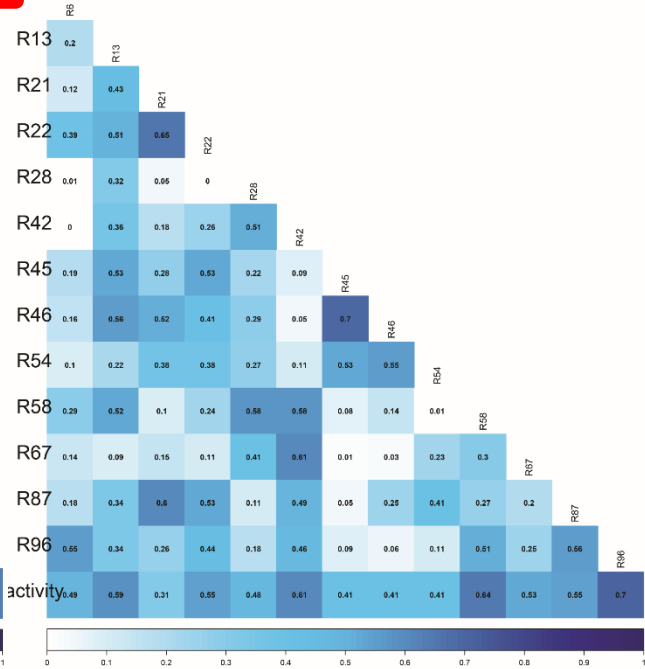
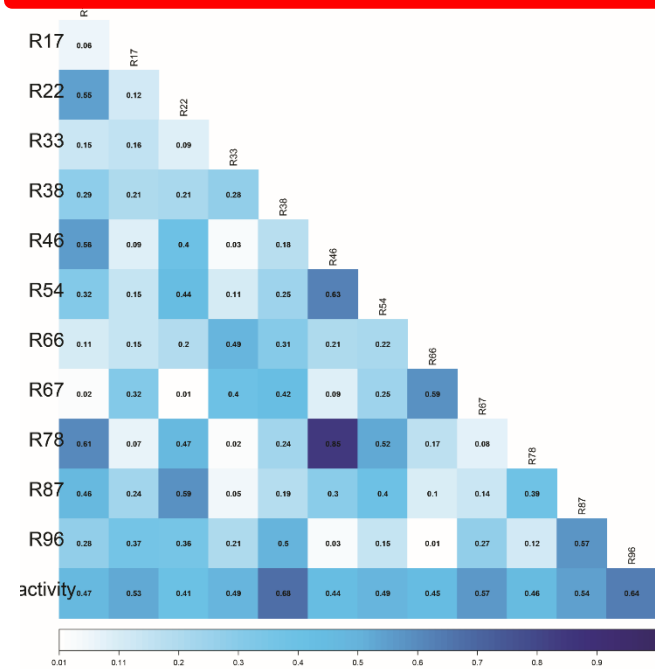
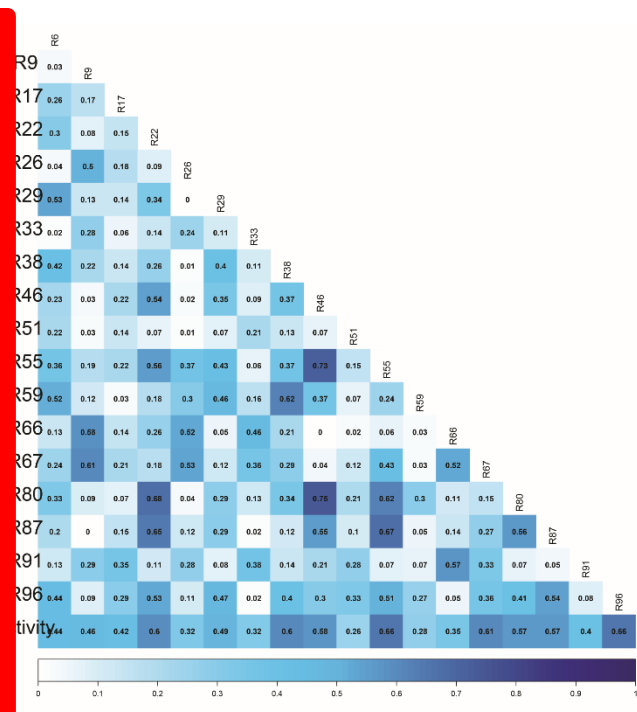
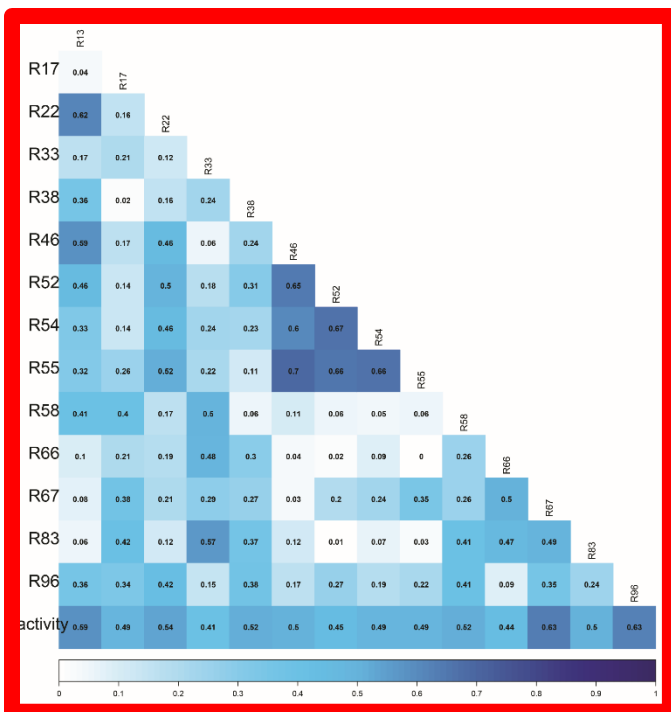
Figure S-3.5: *Residue pairs with greatest anticorrelation*

Residue Pairs with the Greatest Anti-Correlation				
Variant	Variant #	Residue 1	Residue 2	Activity
T91A	91	25	45	1
K20R	20	68	44	1
K55R	55	56	34	1
Q7R	7	51	25	1
L10F	10	25	46	1
L63P	63	22	46	1
N37S	37	65	96	1
R41S	41	85	52	1
M46V	46	42	52	1
I64V	64	99	27	1
A71V	71	57	34	1
I72V	72	99	27	1
V56A	56	51	59	0
T80A	80	33	46	0
G94A	94	50	30	0
T96A	96	95	69	0
Q2R	2	57	79	0
G40R	40	56	79	0
W42R	42	54	76	0
G68R	68	27	52	0
P9R	9	5	72	0
C95R	95	97	70	0
T96N	96	77	53	0
N83D	83	76	79	0
N88D	88	23	59	0
G94D	94	67	57	0
L23Q	23	33	76	0
K20E	20	68	57	0
D25E	25	33	53	0
G27E	27	13	91	0
V32E	32	65	2	0

G40E	40	40	27	0
E21G	21	34	55	0
D25G	25	61	51	0
R57G	57	95	69	0
D60G	60	57	55	0
E65G	65	25	46	0
R87G	87	68	96	0
D29H	29	14	2	0
Y59H	59	34	46	0
P81H	81	20	47	0
T31I	31	25	52	0
I47L	47	82	53	0
I62L	62	50	25	0
V75L	75	82	58	0
P81L	81	56	83	0
E35K	35	42	23	0
I64K	64	65	26	0
R87K	87	70	96	0
I15M	15	25	46	0
I47M	47	93	14	0
I85F	85	83	55	0
L89P	89	14	91	0
L38S	38	83	55	0
G52S	52	24	53	0
L76S	76	14	2	0
R87S	87	14	27	0
N88S	88	12	49	0
L90S	90	9	50	0
I3T	3	51	57	0
I54T	54	52	26	0
I64T	64	8	63	0
I66T	66	35	53	0
I85T	85	56	34	0
C95Y	95	99	99	0
G27V	27	43	81	0
I50V	50	34	45	0
G52V	52	80	57	0

Figure S-3.2: Feature selection from XGBoost models





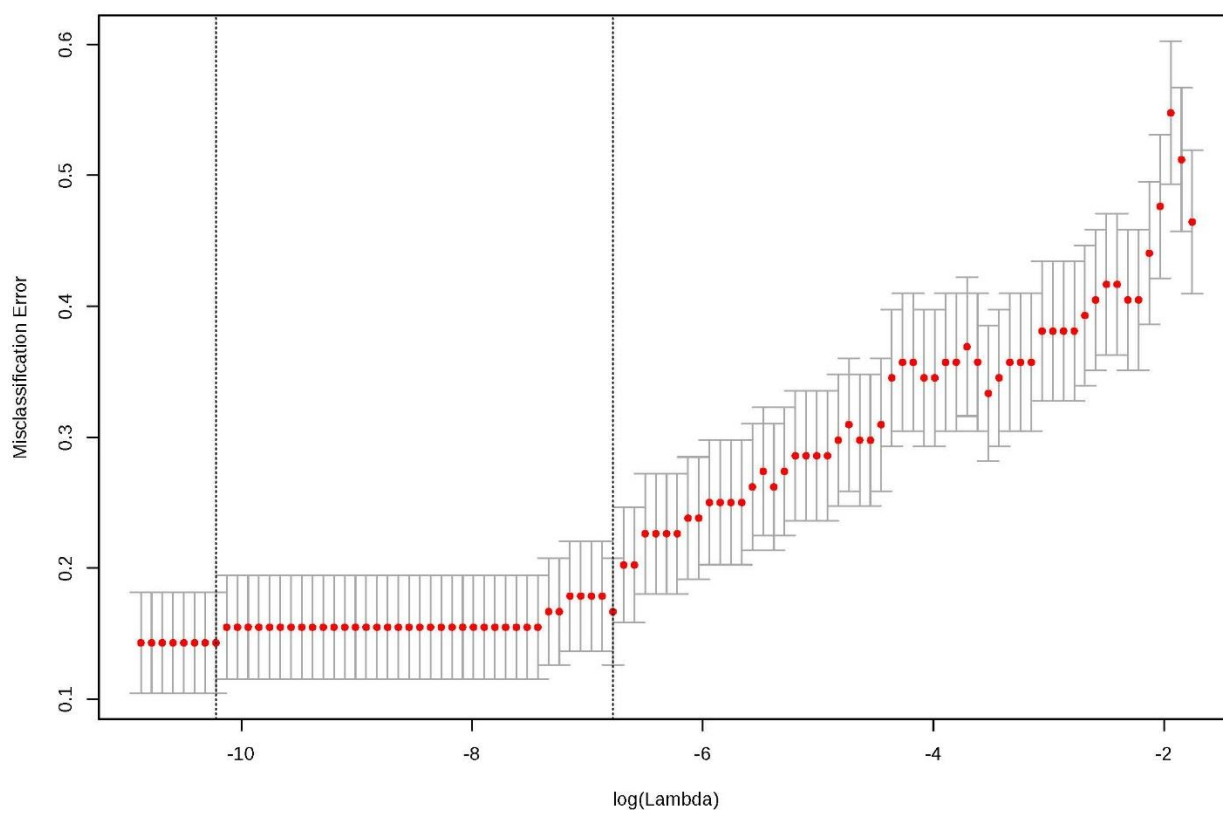


Figure S-3.3: LOOCV Misclassification Error.

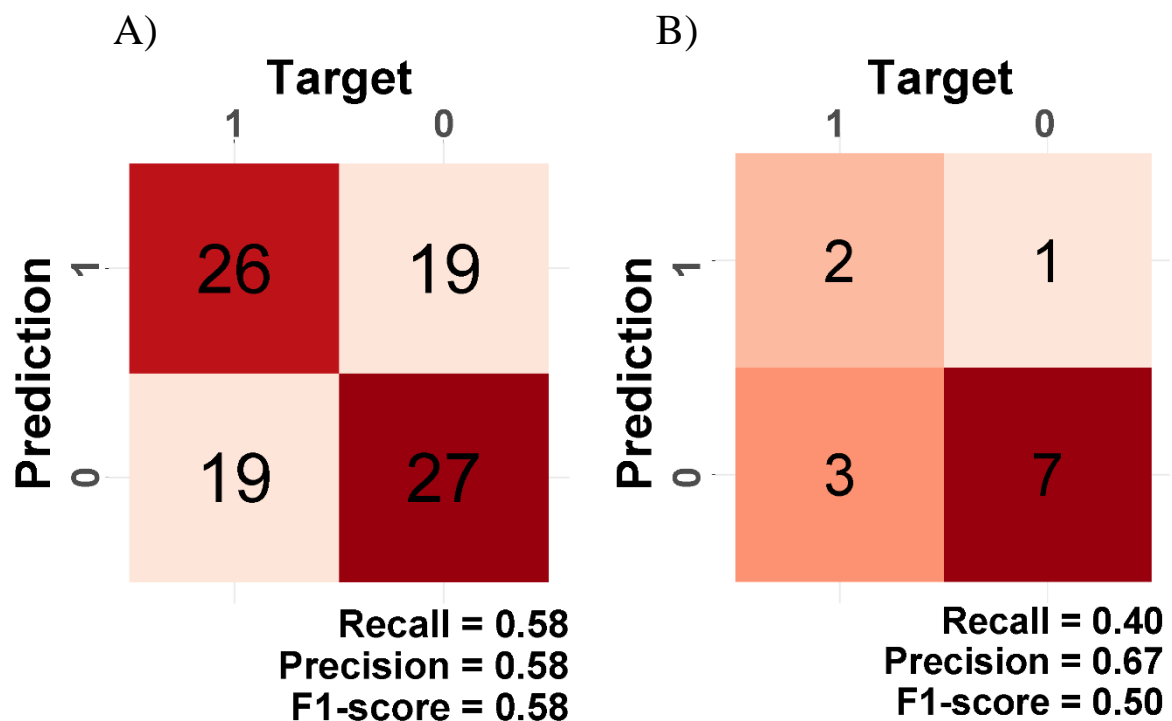


Figure S-3.4: Confusion matrices for the LR model performance **A)** is the performance result of the training set and **B)** is the performance result of the test set with the recall, precision and F1-score shown at the bottom right.

Appendix C

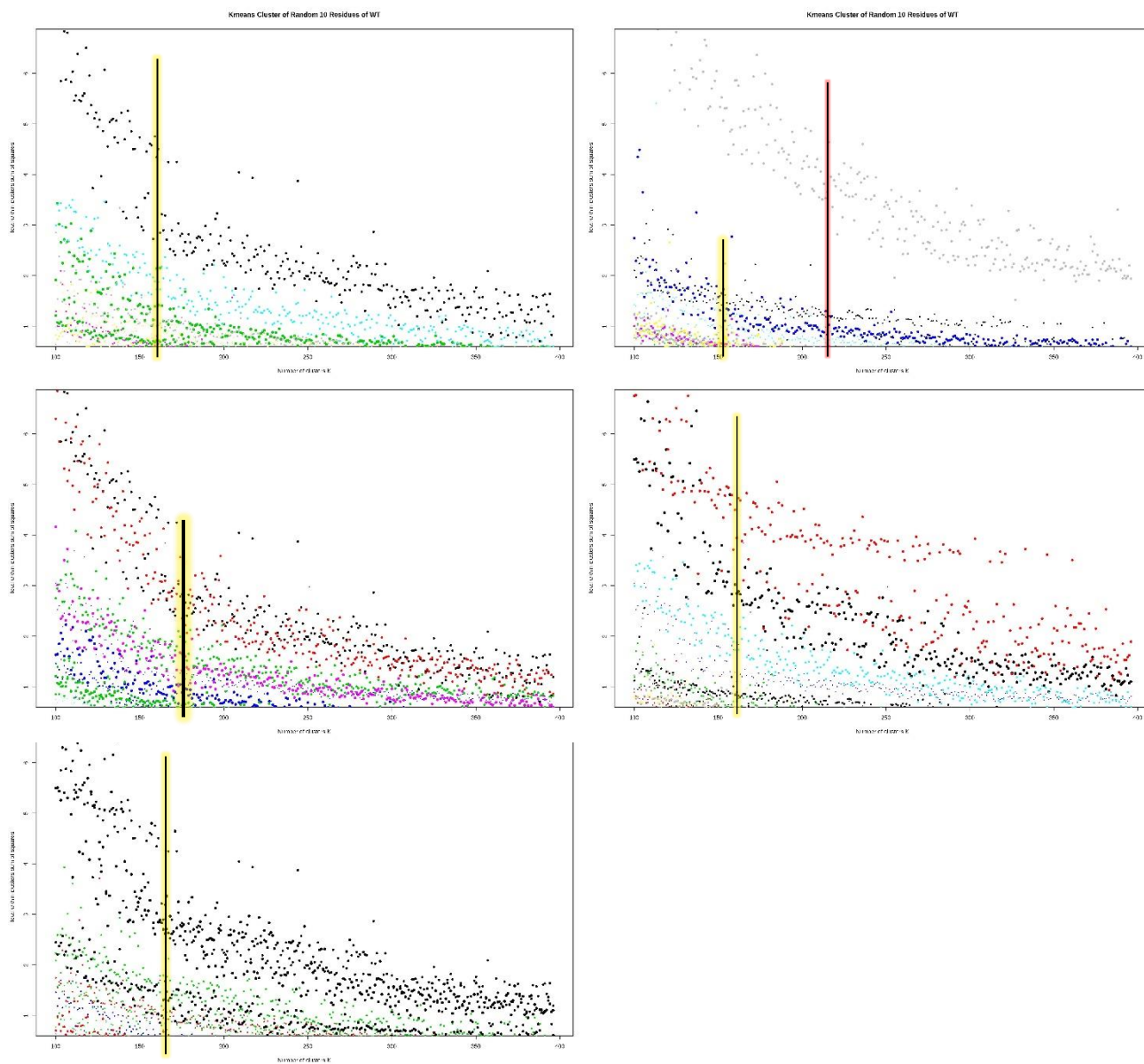
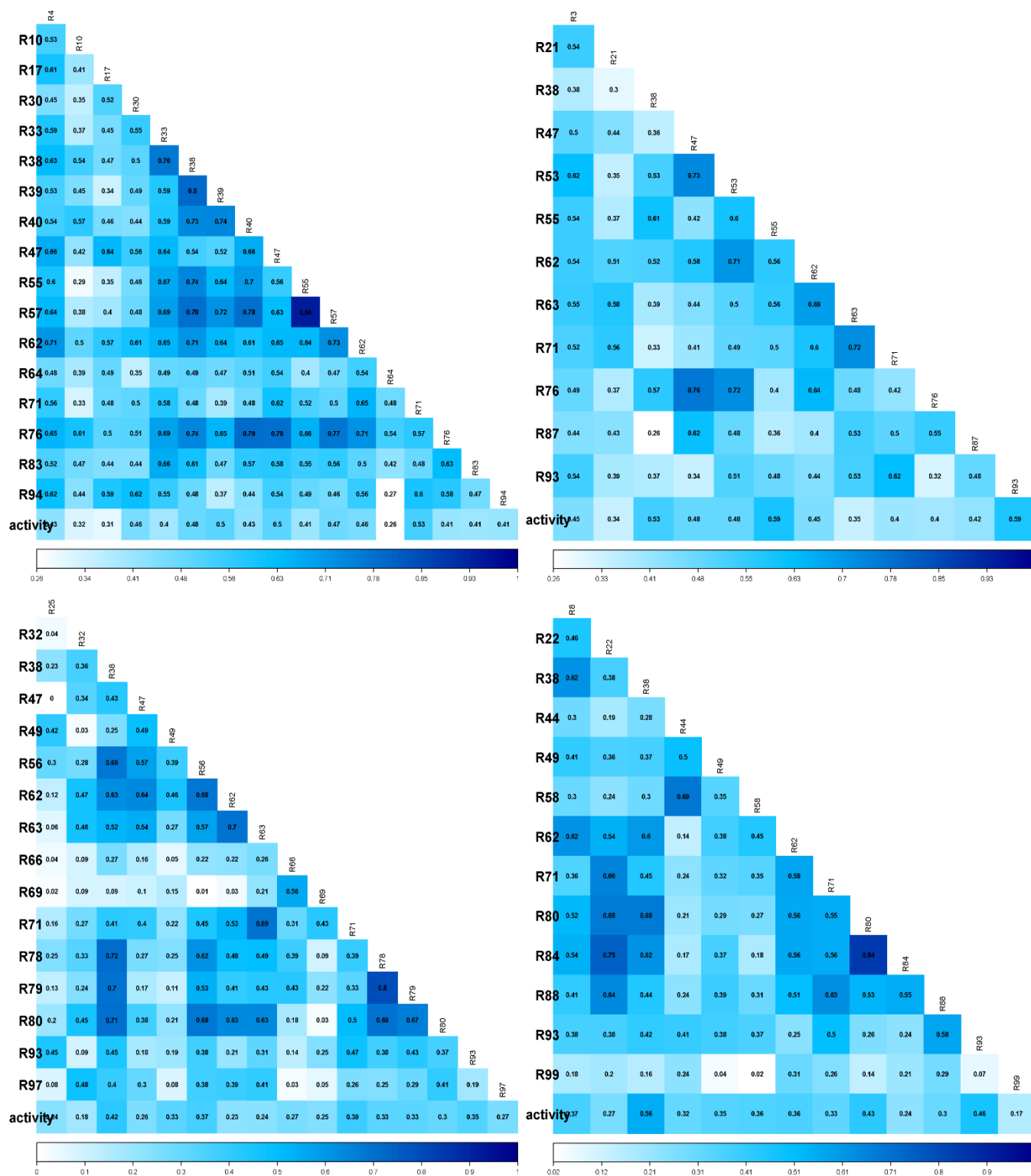
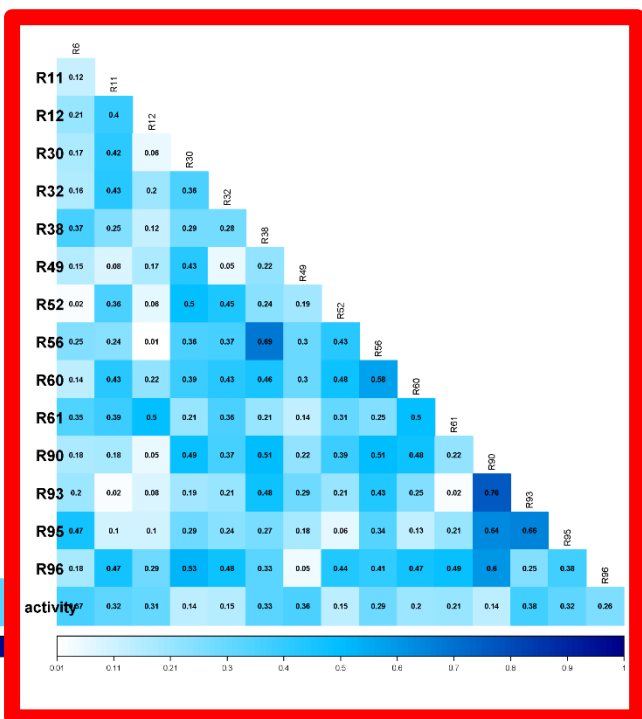
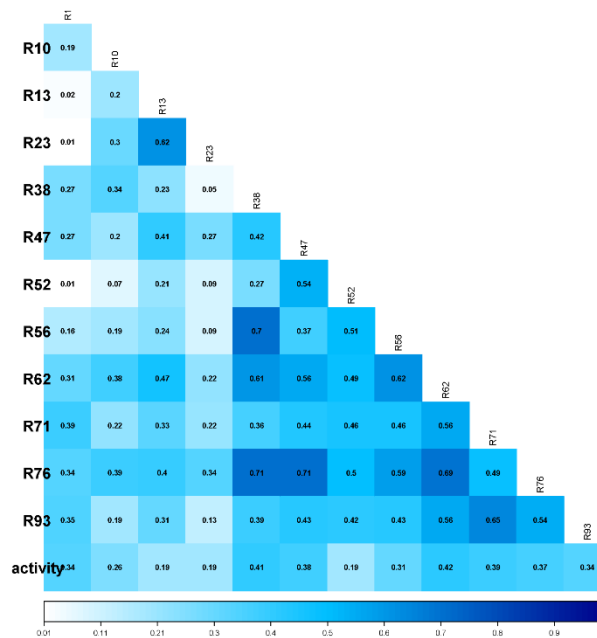
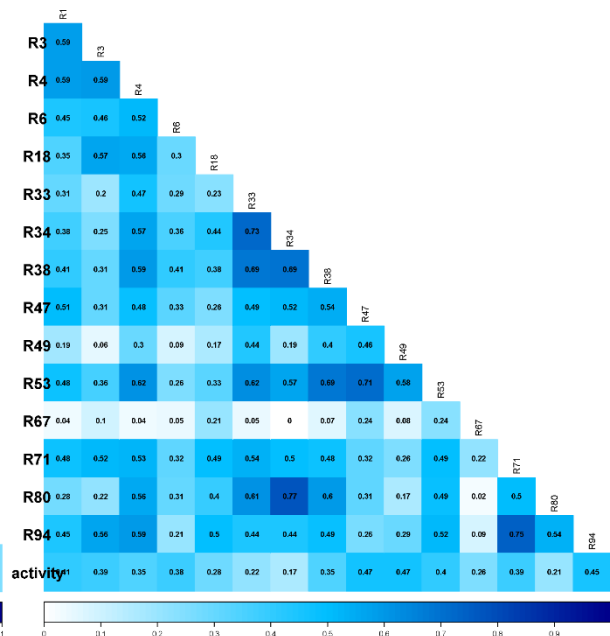
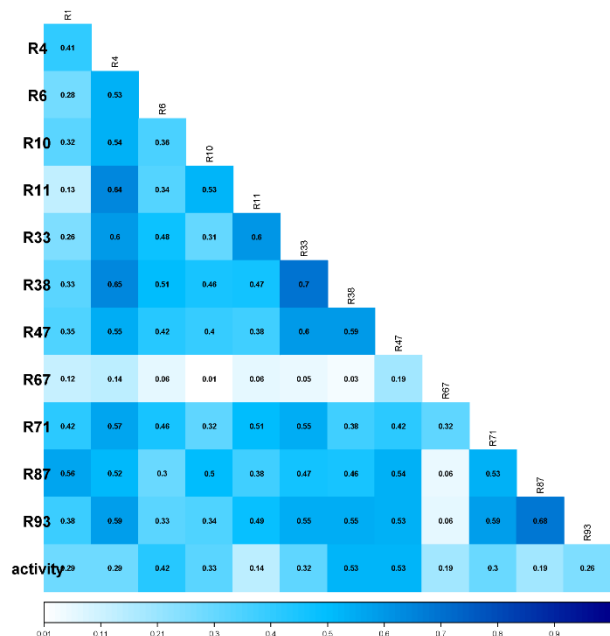


Figure S-4.1: Results from *K*-means algorithm

Figure S-4.2A-H ; Feature selection from the LR model





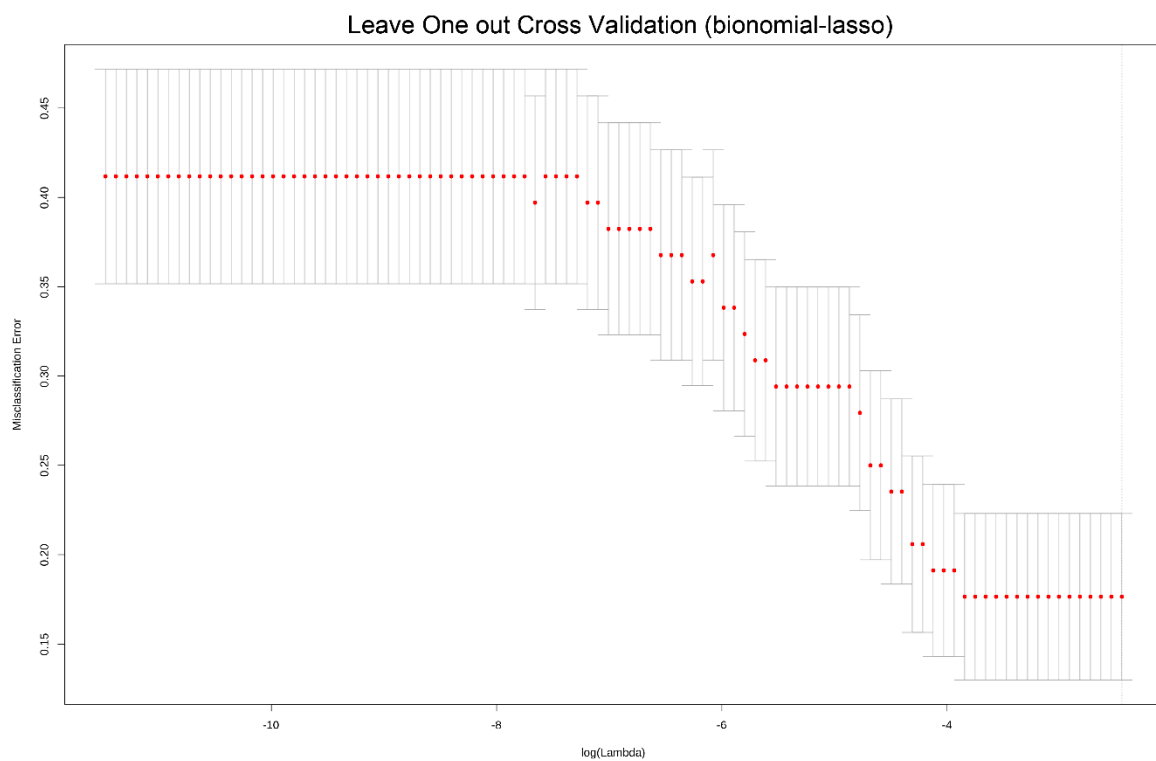


Figure S-4.3: *LOOCV misclassification error ($\log(\text{Lambda})$)*

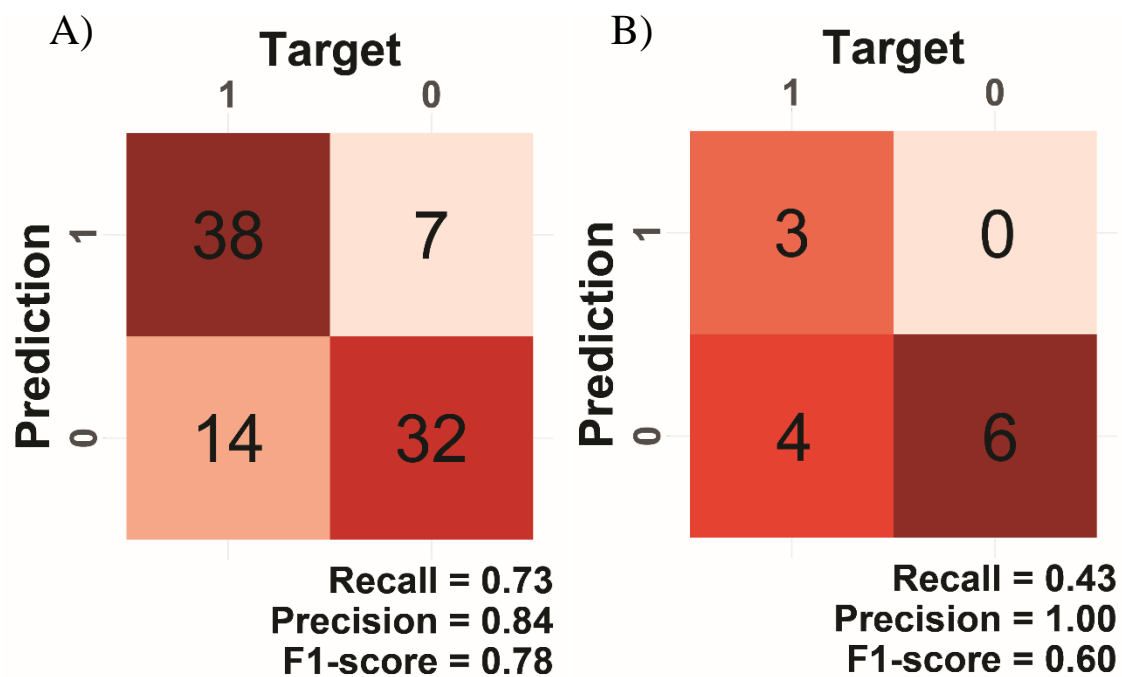


Figure S-4.4: Confusion matrix of the XGBoost model performance **A)** the performance result of training set **B)** the performance result of test set with the recall, precision and F1-score shown at the bottom right.

Figure S-4.1: Total MI for each residue i with the neighboring residues i+1 and i-1.

Total MI for the CAV

Variant	dMir-1	dMir	dMir+1
R41S	4.434	-0.068	-0.689
Q7R	-3.952	-1.599	-0.59
N37S	-0.338	-3.221	-3.265
L10F	0.235	-5.07	-12.118
M46V	-6.27	-7.494	-6.929
K55R	-1.865	0.937	2.816
I64V	2.536	11.474	5.026
L63P	-4.748	2.71	0.474
I72V	-0.384	4.477	5.638
T91A	-0.219	5.731	1.601
A71V	8.931	6.841	5.152
K20R	6.787	7.599	3.408

Total MI for the CIV

Variant	dMir-1	dMir	dMir+1
D60G	-1.734	-0.125	1.486
I85F	1.739	-1.309	-2.101
L89P	2.544	-1.32	1.225
G52S	0.227	-1.522	1.443
P9R	-3.699	-1.619	-6.798
Q2R	4.89	-2.297	3.679
I50V	-12.393	-3.096	0.698
T31I	-1.273	-3.543	6.356
R57G	-0.709	-3.968	-0.823
D25G	-4.258	-3.989	-15.864
N88S	2.27	-4.226	-3.554
E21G	2.543	-4.339	-0.43
D29H	7.42	-4.505	8.543
L76S	4.32	0.327	3.102
Y59H	-0.309	0.359	-2.504
L90S	4.382	0.454	5.162
V75L	1.439	0.48	-5.613
E65G	-1.744	0.52	-0.35
G94D	1.978	0.599	-3.858
I47M	-1.243	0.681	-1.392
G27V	-0.207	0.93	4.013
G94A	-4.453	1.239	-4.334
W42R	2.822	1.274	-0.949
I54T	4.428	1.58	-0.081
N83D	-1.136	1.751	-0.646
T80A	10.919	10.059	12.125
I64K	9.623	10.266	12.284
C95R	10.355	12.2	14.364

Variant	dMir-1	dMir	dMir+1
C95R	10.355	12.2	14.364
R87K	8.344	12.322	9.114
G27E	-4.562	13.327	10.249
P81H	13.195	13.493	8.847
P81L	7.001	14.731	-1.628
T96A	1.56	2.009	2.199
I85T	-1.845	2.182	4.222
G40R	3.389	2.209	1.664
R87S	1.622	2.266	7.672
K20E	1.26	2.597	3.212
I66T	3.003	2.825	0.375
I62L	-3.509	3.65	3.877
V56A	1.501	3.658	1.871
T96N	-1.048	3.701	4.235
G68R	5.272	3.708	8.2
L23Q	1.724	3.908	-2.991
I3T	1.026	3.935	-2.047
V32E	-0.925	4.066	6.191
G52V	-0.23	4.169	3.736
R87G	3.551	4.201	12.276
D25E	7.458	4.324	2.242
I64T	0.26	5.124	4.314
G40E	5.31	5.636	1.017
C95Y	7.247	6.121	5.509
I15M	8.023	6.893	6.783
I47L	1.645	7.074	10.559
L38S	8.424	7.533	9.206
E35K	6.081	7.812	5.41
N88D	5.574	7.909	3.324

Appendix D

Notes from the Challenges and future work

A major concept I looked to cover under chapter I was the difference in internal volume, the empty space within a protein structure. The tools I used to evaluate the empty space were fpocket, which supplies a wealth of detail regarding the volume calculated. However, optimizing the tool to measure each structure equaling requires exhaustive parameterization and ultimately, I had to measure a single monomer collectively rather than identifying a set number of pockets. Given the time spent and the end result and the complexity of the analysis there was little to no supportive evidence. However, I had learned a lot about rationalizing the methods used to compute complex ratios as well as gain a better understanding of aggregation.

Also, while looking into the energies, I had calculated a number of features to include individual energy of a residue. I had done a number of procedures similar to what I had gone through with looking at the internal volume. Ultimately the scale and computational load was extremely high and a constant battle, but I believe there is some details within the changes in a residues internal energy that can help understand allostery more.

Future work, I'm creative and have goals in mind and I do see the expansive directions things can go. Focusing solely on the study of protein dynamics it would be interesting to study variant structures that have no known activity using the model I had

developed as well as work towards understanding the application of other methods and datasets.

As I have started a career before completing this, the group I work with studies exposure science, whereas the research that I have begun there has broaden my knowledge of the resources available and the scope of how I will institute some of these concepts moving forward.

REFERENCES

- (1) Parera, M.; Fernandez, G.; Clotet, B.; Martinez, M. A. HIV-1 protease catalytic efficiency effects caused by random single amino acid substitutions. *Mol Biol Evol* **2007**, *24* (2), 382-387. DOI: 10.1093/molbev/msl168.
- (2) Caldararu, O.; Kumar, R.; Oksanen, E.; Logan, D. T.; Ryde, U. Are crystallographic B-factors suitable for calculating protein conformational entropy? *Phys Chem Chem Phys* **2019**, *21* (33), 18149-18160. DOI: 10.1039/c9cp02504a.
- (3) Zhang, C.; Samad, M.; Yu, H.; Chakraborty, N.; Hilton, D.; Dalby, P. A. Computational Design To Reduce Conformational Flexibility and Aggregation Rates of an Antibody Fab Fragment. *Mol Pharm* **2018**, *15* (8), 3079-3092. DOI: 10.1021/acs.molpharmaceut.8b00186.
- (4) Christensen, N. J.; Kepp, K. P. Stability mechanisms of a thermophilic laccase probed by molecular dynamics. *PLoS One* **2013**, *8* (4), e61985. DOI: 10.1371/journal.pone.0061985.
- (5) Wang, Z.; Abeysinghe, T.; Finer-Moore, J. S.; Stroud, R. M.; Kohen, A. A remote mutation affects the hydride transfer by disrupting concerted protein motions in thymidylate synthase. *J Am Chem Soc* **2012**, *134* (42), 17722-17730. DOI: 10.1021/ja307859m.
- (6) Catazaro, J.; Caprez, A.; Swanson, D.; Powers, R. Functional Evolution of Proteins. *Proteins* **2019**, *87* (6), 492-501. DOI: 10.1002/prot.25670.
- (7) Forbes, J.; Krishnamurthy, K. Biochemistry, Peptide. In *StatPearls*, 2022.
- (8) Meringer, M.; Cleaves, H. J., 2nd; Freeland, S. J. Beyond terrestrial biology: charting the chemical universe of alpha-amino acid structures. *J Chem Inf Model* **2013**, *53* (11), 2851-2862. DOI: 10.1021/ci400209n.
- (9) Takahashi, T.; Vo Ngo, B. C.; Xiao, L.; Arya, G.; Heller, M. J. Molecular mechanical properties of short-sequence peptide enzyme mimics. *J Biomol Struct Dyn* **2016**, *34* (3), 463-474. DOI: 10.1080/07391102.2015.1039586.
- (10) Zardecki, C.; Dutta, S.; Goodsell, D. S.; Lowe, R.; Voigt, M.; Burley, S. K. PDB-101: Educational resources supporting molecular explorations through biology and medicine. *Protein Sci* **2022**, *31* (1), 129-140. DOI: 10.1002/pro.4200.
- (11) Samanta, S.; Mukherjee, S. Co-operative intra-protein structural response due to protein-protein complexation revealed through thermodynamic quantification: study of MDM2-p53 binding. *J Comput Aided Mol Des* **2017**, *31* (10), 891-903. DOI: 10.1007/s10822-017-0057-y.
- (12) Abel, S.; Dupradeau, F. Y.; Marchi, M. Molecular Dynamics Simulations of a Characteristic DPC Micelle in Water. *J Chem Theory Comput* **2012**, *8* (11), 4610-4623. DOI: 10.1021/ct3003207.
- (13) Buck, P. M.; Bystroff, C. Simulating protein folding initiation sites using an alpha-carbon-only knowledge-based force field. *Proteins* **2009**, *76* (2), 331-342. DOI: 10.1002/prot.22348.
- (14) Feyfant, E.; Sali, A.; Fiser, A. Modeling mutations in protein structures. *Protein Sci* **2007**, *16* (9), 2030-2041. DOI: 10.1110/ps.072855507.
- (15) Bedard, L.; Lefevre, T.; Morin-Michaud, E.; Auger, M. Besides fibrillization: putative role of the peptide fragment 71-82 on the structural and assembly behavior of alpha-synuclein. *Biochemistry* **2014**, *53* (41), 6463-6472. DOI: 10.1021/bi5008707.
- (16) Hendus-Altenburger, R.; Fernandes, C. B.; Bugge, K.; Kunze, M. B. A.; Boomsma, W.; Kragelund, B. B. Random coil chemical shifts for serine, threonine and tyrosine phosphorylation over a broad pH range. *J Biomol NMR* **2019**, *73* (12), 713-725. DOI: 10.1007/s10858-019-00283-z.
- (17) Chu, W. T.; Nesbitt, N. M.; Gnatenko, D. V.; Li, Z.; Zhang, B.; Seeliger, M. A.; Browne, S.; Mantle, T. J.; Bahou, W. F.; Wang, J. Enzymatic Activity and Thermodynamic Stability of Biliverdin IXbeta Reductase Are Maintained by an Active Site Serine. *Chemistry* **2017**, *23* (8), 1891-1900. DOI: 10.1002/chem.201604517.
- (18) von Grafenstein, S.; Wallnoefer, H. G.; Kirchmair, J.; Fuchs, J. E.; Huber, R. G.; Schmidtke, M.; Sauerbrei, A.; Rollinger, J. M.; Liedl, K. R. Interface dynamics explain assembly dependency of influenza neuraminidase catalytic activity. *J Biomol Struct Dyn* **2015**, *33* (1), 104-120. DOI: 10.1080/07391102.2013.855142.

- (19) Shelar, A.; Bansal, M. Sequence and conformational preferences at termini of alpha-helices in membrane proteins: role of the helix environment. *Proteins* **2014**, *82* (12), 3420-3436. DOI: 10.1002/prot.24696.
- (20) Wang, X. Y.; Ji, C. G.; Zhang, J. Z. Exploring the molecular mechanism of stabilization of the adhesion domains of human CD2 by N-glycosylation. *J Phys Chem B* **2012**, *116* (38), 11570-11577. DOI: 10.1021/jp304116d.
- (21) Li, X.; Dong, C.; Hoffmann, M.; Garen, C. R.; Cortez, L. M.; Petersen, N. O.; Woodside, M. T. Early stages of aggregation of engineered alpha-synuclein monomers and oligomers in solution. *Sci Rep* **2019**, *9* (1), 1734. DOI: 10.1038/s41598-018-37584-6.
- (22) Kulesza, A.; Marklund, E. G.; MacAleese, L.; Chirot, F.; Dugourd, P. Bringing Molecular Dynamics and Ion-Mobility Spectrometry Closer Together: Shape Correlations, Structure-Based Predictors, and Dissociation. *J Phys Chem B* **2018**, *122* (35), 8317-8329. DOI: 10.1021/acs.jpcc.8b03825.
- (23) Eginton, C.; Cressman, W. J.; Bachas, S.; Wade, H.; Beckett, D. Allosteric coupling via distant disorder-to-order transitions. *J Mol Biol* **2015**, *427* (8), 1695-1704. DOI: 10.1016/j.jmb.2015.02.021.
- (24) Sawyer, N.; Watkins, A. M.; Arora, P. S. Protein Domain Mimics as Modulators of Protein-Protein Interactions. *Acc Chem Res* **2017**, *50* (6), 1313-1322. DOI: 10.1021/acs.accounts.7b00130.
- (25) Rose, P. W.; Bi, C.; Bluhm, W. F.; Christie, C. H.; Dimitropoulos, D.; Dutta, S.; Green, R. K.; Goodsell, D. S.; Prlic, A.; Quesada, M.; et al. The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res* **2013**, *41* (Database issue), D475-482. DOI: 10.1093/nar/gks1200.
- (26) Berman, H.; Henrick, K.; Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* **2003**, *10* (12), 980. DOI: 10.1038/nsb1203-980.
- (27) Kundrotas, P. J.; Vakser, I. A.; Janin, J. Structural templates for modeling homodimers. *Protein Sci* **2013**, *22* (11), 1655-1663. DOI: 10.1002/pro.2361.
- (28) Bahar, I.; Lezon, T. R.; Yang, L. W.; Eyal, E. Global Dynamics of Proteins: Bridging Between Structure and Function. *Annual Review of Biophysics*, Vol 39 **2010**, *39*, 23-42. DOI: 10.1146/annurev.biophys.093008.131258.
- (29) Funatsu, M. *Proteins: structure and function*; Kodansha; Wiley, 1972.
- (30) van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R. J.; Daughdrill, G. W.; Dunker, A. K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D. T.; et al. Classification of Intrinsically Disordered Regions and Proteins. *Chem Rev* **2014**, *114* (13), 6589-6631. DOI: 10.1021/cr400525m.
- (31) Babu, M. M.; van der Lee, R.; de Groot, N. S.; Gsponer, J. Intrinsically disordered proteins: regulation and disease. *Curr Opin Struct Biol* **2011**, *21* (3), 432-440. DOI: 10.1016/j.sbi.2011.03.011.
- (32) Szklarczyk, D.; Morris, J. H.; Cook, H.; Kuhn, M.; Wyder, S.; Simonovic, M.; Santos, A.; Doncheva, N. T.; Roth, A.; Bork, P.; et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* **2017**, *45* (D1), D362-D368. DOI: 10.1093/nar/gkw937.
- (33) Neves, R. P. P.; Fernandes, P. A.; Ramos, M. J. Role of Enzyme and Active Site Conformational Dynamics in the Catalysis by alpha-Amylase Explored with QM/MM Molecular Dynamics. *J Chem Inf Model* **2022**, *62* (15), 3638-3650. DOI: 10.1021/acs.jcim.2c00691.
- (34) Ahsan, M.; Pindi, C.; Senapati, S. Electrostatics Plays a Crucial Role in HIV-1 Protease Substrate Binding, Drugs Fail to Take Advantage. *Biochemistry* **2020**, *59* (36), 3316-3331. DOI: 10.1021/acs.biochem.0c00341.
- (35) Wang, Y.; V, S. M.; Kim, J.; Li, G.; Ahuja, L. G.; Aoto, P.; Taylor, S. S.; Veglia, G. Globally correlated conformational entropy underlies positive and negative cooperativity in a kinase's enzymatic cycle. *Nat Commun* **2019**, *10* (1), 799. DOI: 10.1038/s41467-019-08655-7.
- (36) Ahuja, L. G.; Aoto, P. C.; Kornev, A. P.; Veglia, G.; Taylor, S. S. Dynamic allostery-based molecular workings of kinase:peptide complexes. *Proc Natl Acad Sci U S A* **2019**, *116* (30), 15052-15061. DOI: 10.1073/pnas.1900163116.
- (37) Maria-Solano, M. A.; Serrano-Hervas, E.; Romero-Rivera, A.; Iglesias-Fernandez, J.; Osuna, S. Role of conformational dynamics in the evolution of novel enzyme function. *Chem Commun (Camb)* **2018**, *54* (50), 6622-6634. DOI: 10.1039/c8cc02426j.
- (38) Schupfner, M.; Straub, K.; Busch, F.; Merkl, R.; Sterner, R. Analysis of allosteric communication in a multienzyme complex by ancestral sequence reconstruction. *Proc Natl Acad Sci U S A* **2020**, *117* (1), 346-354. DOI: 10.1073/pnas.1912132117.

- (39) La Sala, G.; Decherchi, S.; De Vivo, M.; Rocchia, W. Allosteric Communication Networks in Proteins Revealed through Pocket Crosstalk Analysis. *ACS Cent Sci* **2017**, *3* (9), 949-960. DOI: 10.1021/acscentsci.7b00211.
- (40) Hoyte, A. C.; Jamin, A. V.; Koneru, P. C.; Kobe, M. J.; Larue, R. C.; Fuchs, J. R.; Engelman, A. N.; Kvaratskhelia, M. Resistance to pyridine-based inhibitor KF116 reveals an unexpected role of integrase in HIV-1 Gag-Pol polyprotein proteolytic processing. *J Biol Chem* **2017**, *292* (48), 19814-19825. DOI: 10.1074/jbc.M117.816645.
- (41) Mastorodemos, V.; Kanavouras, K.; Sundaram, S.; Providaki, M.; Petraki, Z.; Kokkinidis, M.; Zaganas, I.; Logothetis, D. E.; Plaitakis, A. Side-chain interactions in the regulatory domain of human glutamate dehydrogenase determine basal activity and regulation. *J Neurochem* **2015**, *133* (1), 73-82. DOI: 10.1111/jnc.13019.
- (42) Nussinov, R. Allosteric modulators can restore function in an amino acid neurotransmitter receptor by slightly altering intra-molecular communication pathways. *Br J Pharmacol* **2012**, *165* (7), 2110-2112. DOI: 10.1111/j.1476-5381.2011.01793.x.
- (43) Schweizer, R. S.; Aponte, R. A.; Zimmermann, S.; Weber, A.; Reinstein, J. Fine tuning of a biological machine: DnaK gains improved chaperone activity by altered allosteric communication and substrate binding. *Chembiochem* **2011**, *12* (10), 1559-1573. DOI: 10.1002/cbic.201000786.
- (44) Semmelmann, F.; Straub, K.; Nazet, J.; Rajendran, C.; Merkl, R.; Sterner, R. Mapping the Allosteric Communication Network of Aminodeoxychorismate Synthase. *J Mol Biol* **2019**. DOI: 10.1016/j.jmb.2019.05.021.
- (45) Schulze, J. O.; Saladino, G.; Busschots, K.; Neimanis, S.; Suss, E.; Odadzic, D.; Zeuzem, S.; Hindie, V.; Herbrand, A. K.; Lisa, M. N.; et al. Bidirectional Allosteric Communication between the ATP-Binding Site and the Regulatory PIF Pocket in PDK1 Protein Kinase. *Cell Chem Biol* **2016**, *23* (10), 1193-1205. DOI: 10.1016/j.chembiol.2016.06.017.
- (46) Guarnera, E.; Berezovsky, I. N. Structure-Based Statistical Mechanical Model Accounts for the Causality and Energetics of Allosteric Communication. *PLoS Comput Biol* **2016**, *12* (3), e1004678. DOI: 10.1371/journal.pcbi.1004678.
- (47) Allain, A.; Chauvot de Beauchene, I.; Langenfeld, F.; Guarracino, Y.; Laine, E.; Tchertanov, L. Allosteric pathway identification through network analysis: from molecular dynamics simulations to interactive 2D and 3D graphs. *Faraday Discuss* **2014**, *169*, 303-321. DOI: 10.1039/c4fd00024b.
- (48) VanSchouwen, B.; Selvaratnam, R.; Fogolari, F.; Melacini, G. Role of dynamics in the autoinhibition and activation of the exchange protein directly activated by cyclic AMP (EPAC). *J Biol Chem* **2011**, *286* (49), 42655-42669. DOI: 10.1074/jbc.M111.277723.
- (49) Bhattacharyya, M.; Ghosh, A.; Hansia, P.; Vishveshwara, S. Allostery and conformational free energy changes in human tryptophanyl-tRNA synthetase from essential dynamics and structure networks. *Proteins* **2010**, *78* (3), 506-517. DOI: 10.1002/prot.22573.
- (50) Goodey, N. M.; Benkovic, S. J. Allosteric regulation and catalysis emerge via a common route. *Nat Chem Biol* **2008**, *4* (8), 474-482. DOI: 10.1038/nchembio.98.
- (51) Gerhart, J. Enzymes, embryos, and ancestors. *Annu Rev Cell Dev Biol* **2010**, *26*, 1-20. DOI: 10.1146/annurev-cellbio-100109-104101.
- (52) Guo, J.; Zhou, H. X. Dynamically Driven Protein Allostery Exhibits Disparate Responses for Fast and Slow Motions. *Biophys J* **2015**, *108* (12), 2771-2774. DOI: 10.1016/j.bpj.2015.04.035.
- (53) Motlagh, H. N.; Wrabl, J. O.; Li, J.; Hilser, V. J. The ensemble nature of allostery. *Nature* **2014**, *508* (7496), 331-339. DOI: 10.1038/nature13001.
- (54) Gunasekaran, K.; Ma, B.; Nussinov, R. Is allostery an intrinsic property of all dynamic proteins? *Proteins: Structure, Function, and Bioinformatics* **2004**, *57* (3), 433-443. DOI: 10.1002/prot.20232 (accessed 2020/03/09).
- (55) Burley, S. K.; Berman, H. M.; Christie, C.; Duarte, J. M.; Feng, Z.; Westbrook, J.; Young, J.; Zardecki, C. RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Sci* **2018**, *27* (1), 316-330. DOI: 10.1002/pro.3331.
- (56) Burley, S. K.; Berman, H. M.; Bhikadiya, C.; Bi, C.; Chen, L.; Di Costanzo, L.; Christie, C.; Dalenberg, K.; Duarte, J. M.; Dutta, S.; et al. RCSB Protein Data Bank: biological macromolecular structures enabling

research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res* **2019**, *47* (D1), D464-D474. DOI: 10.1093/nar/gky1004.

(57) Goodsell, D. S.; Zardecki, C.; Di Costanzo, L.; Duarte, J. M.; Hudson, B. P.; Persikova, I.; Segura, J.; Shao, C.; Voigt, M.; Westbrook, J. D.; et al. RCSB Protein Data Bank: Enabling biomedical research and drug discovery. *Protein Sci* **2020**, *29* (1), 52-65. DOI: 10.1002/pro.3730.

(58) Burley, S. K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chen, L.; Crichlow, G. V.; Christie, C. H.; Dalenberg, K.; Di Costanzo, L.; Duarte, J. M.; et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res* **2021**, *49* (D1), D437-D451. DOI: 10.1093/nar/gkaa1038.

(59) Bittrich, S.; Rose, Y.; Segura, J.; Lowe, R.; Westbrook, J. D.; Duarte, J. M.; Burley, S. K. RCSB Protein Data Bank: improved annotation, search and visualization of membrane protein structures archived in the PDB. *Bioinformatics* **2022**, *38* (5), 1452-1454. DOI: 10.1093/bioinformatics/btab813.

(60) Burley, S. K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chao, H.; Chen, L.; Craig, P. A.; Crichlow, G. V.; Dalenberg, K.; Duarte, J. M.; et al. RCSB Protein Data bank: Tools for visualizing and understanding biological macromolecules in 3D. *Protein Sci* **2022**, *31* (12), e4482. DOI: 10.1002/pro.4482.

(61) Duarte, J. M.; Dutta, S.; Goodsell, D. S.; Burley, S. K. Exploring protein symmetry at the RCSB Protein Data Bank. *Emerg Top Life Sci* **2022**, *6* (3), 231-243. DOI: 10.1042/ETLS20210267.

(62) Burley, S. K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chao, H.; Chen, L.; Craig, P. A.; Crichlow, G. V.; Dalenberg, K.; Duarte, J. M.; et al. RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res* **2023**, *51* (D1), D488-D508. DOI: 10.1093/nar/gkac1077.

(63) Zoi, I.; Suarez, J.; Antoniou, D.; Cameron, S. A.; Schramm, V. L.; Schwartz, S. D. Modulating Enzyme Catalysis through Mutations Designed to Alter Rapid Protein Dynamics. *J Am Chem Soc* **2016**, *138* (10), 3403-3409. DOI: 10.1021/jacs.5b12551.

(64) Kohen, A. Role of dynamics in enzyme catalysis: substantial versus semantic controversies. *Acc Chem Res* **2015**, *48* (2), 466-473. DOI: 10.1021/ar500322s.

(65) Klinman, J. P. Dynamically achieved active site precision in enzyme catalysis. *Acc Chem Res* **2015**, *48* (2), 449-456. DOI: 10.1021/ar5003347.

(66) Klinman, J. P. Importance of protein dynamics during enzymatic C-H bond cleavage catalysis. *Biochemistry* **2013**, *52* (12), 2068-2077. DOI: 10.1021/bi301504m.

(67) Aden, J.; Verma, A.; Schug, A.; Wolf-Watz, M. Modulation of a pre-existing conformational equilibrium tunes adenylate kinase activity. *J Am Chem Soc* **2012**, *134* (40), 16562-16570. DOI: 10.1021/ja3032482.

(68) Kamerlin, S. C.; Warshel, A. At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis? *Proteins* **2010**, *78* (6), 1339-1375. DOI: 10.1002/prot.22654.

(69) Radkiewicz, J. L.; Brooks, C. L. Protein dynamics in enzymatic catalysis: Exploration of dihydrofolate reductase. *Journal of the American Chemical Society* **2000**, *122* (2), 225-231. DOI: DOI 10.1021/ja9913838.

(70) Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Henin, J.; Jiang, W.; et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J Chem Phys* **2020**, *153* (4), 044130. DOI: 10.1063/5.0014475.

(71) Jo, S.; Cheng, X.; Lee, J.; Kim, S.; Park, S. J.; Patel, D. S.; Beaven, A. H.; Lee, K. I.; Rui, H.; Park, S.; et al. CHARMM-GUI 10 years for biomolecular modeling and simulation. *J Comput Chem* **2017**, *38* (15), 1114-1124. DOI: 10.1002/jcc.24660.

(72) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J Comput Chem* **2005**, *26* (16), 1781-1802. DOI: 10.1002/jcc.20289.

(73) Cooper, A.; Dryden, D. T. Allostery without conformational change. A plausible model. *Eur Biophys J* **1984**, *11* (2), 103-109. DOI: 10.1007/BF00276625.

(74) Kern, D.; Zuiderweg, E. R. The role of dynamics in allosteric regulation. *Curr Opin Struct Biol* **2003**, *13* (6), 748-757.

- (75) McClendon, C. L.; Friedland, G.; Mobley, D. L.; Amirkhani, H.; Jacobson, M. P. Quantifying Correlations Between Allosteric Sites in Thermodynamic Ensembles. *J Chem Theory Comput* **2009**, *5* (9), 2486-2502. DOI: 10.1021/ct9001812.
- (76) Daily, M. D.; Gray, J. J. Allosteric communication occurs via networks of tertiary and quaternary motions in proteins. *PLoS Comput Biol* **2009**, *5* (2), e1000293. DOI: 10.1371/journal.pcbi.1000293.
- (77) Veglia, G.; Cembran, A. Role of conformational entropy in the activity and regulation of the catalytic subunit of protein kinase A. *FEBS J* **2013**, *280* (22), 5608-5615. DOI: 10.1111/febs.12462.
- (78) Zhang, Y.; Chen, H. F. Allosteric mechanism of an oximino-piperidino-piperidine antagonist for the CCR5 chemokine receptor. *Chem Biol Drug Des* **2020**, *95* (1), 113-123. DOI: 10.1111/cbdd.13627.
- (79) Ganguly, D.; Chen, J. Atomistic details of the disordered states of KID and pKID. Implications in coupled binding and folding. *J Am Chem Soc* **2009**, *131* (14), 5214-5223. DOI: 10.1021/ja808999m.
- (80) Li, X.; Hassan, S. A.; Mehler, E. L. Long dynamics simulations of proteins using atomistic force fields and a continuum representation of solvent effects: calculation of structural and dynamic properties. *Proteins* **2005**, *60* (3), 464-484. DOI: 10.1002/prot.20470.
- (81) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chem Rev* **2016**, *116* (14), 7898-7936. DOI: 10.1021/acs.chemrev.6b00163.
- (82) Riniker, S.; Allison, J. R.; van Gunsteren, W. F. On developing coarse-grained models for biomolecular simulation: a review. *Phys Chem Chem Phys* **2012**, *14* (36), 12423-12430. DOI: 10.1039/c2cp40934h.
- (83) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **2010**, *52* (1), 99-105. DOI: 10.1016/j.ymeth.2010.06.002.
- (84) Holliday, M. J.; Camilloni, C.; Armstrong, G. S.; Vendruscolo, M.; Eisenmesser, E. Z. Networks of Dynamic Allostery Regulate Enzyme Function. *Structure* **2017**, *25* (2), 276-286. DOI: 10.1016/j.str.2016.12.003.
- (85) Hacisuleyman, A.; Erman, B. Entropy Transfer between Residue Pairs and Allostery in Proteins: Quantifying Allosteric Communication in Ubiquitin. *PLoS Comput Biol* **2017**, *13* (1), e1005319. DOI: 10.1371/journal.pcbi.1005319.
- (86) Chen, J.; Dima, R. I.; Thirumalai, D. Allosteric communication in dihydrofolate reductase: signaling network and pathways for closed to occluded transition and back. *J Mol Biol* **2007**, *374* (1), 250-266. DOI: 10.1016/j.jmb.2007.08.047.
- (87) Verkhivker, G. M. Network-based modelling and percolation analysis of conformational dynamics and activation in the CDK2 and CDK4 proteins: dynamic and energetic polarization of the kinase lobes may determine divergence of the regulatory mechanisms. *Mol Biosyst* **2017**, *13* (11), 2235-2253. DOI: 10.1039/c7mb00355b.
- (88) Isaksen, G. V.; Aqvist, J.; Brandsdal, B. O. Enzyme surface rigidity tunes the temperature dependence of catalytic rates. *Proc Natl Acad Sci U S A* **2016**, *113* (28), 7822-7827. DOI: 10.1073/pnas.1605237113.
- (89) Karshikoff, A.; Nilsson, L.; Ladenstein, R. Rigidity versus flexibility: the dilemma of understanding protein thermal stability. *FEBS J* **2015**, *282* (20), 3899-3917. DOI: 10.1111/febs.13343.
- (90) Jiang, Y.; Kirmizialtin, S.; Sanchez, I. C. Dynamic void distribution in myoglobin and five mutants. *Sci Rep* **2014**, *4*, 4011. DOI: 10.1038/srep04011.
- (91) Brady, G. P., Jr.; Stouten, P. F. Fast prediction and visualization of protein binding pockets with PASS. *J Comput Aided Mol Des* **2000**, *14* (4), 383-401. DOI: 10.1023/a:1008124202956.
- (92) Dellisanti, C. D.; Hanson, S. M.; Chen, L.; Czajkowski, C. Packing of the Extracellular Domain Hydrophobic Core Has Evolved to Facilitate Pentameric Ligand-gated Ion Channel Function. *Journal of Biological Chemistry* **2011**, *286* (5), 3658-3670. DOI: 10.1074/jbc.M110.156851.
- (93) Guo, R.; Cang, Z.; Yao, J.; Kim, M.; Deans, E.; Wei, G.; Kang, S. G.; Hong, H. Structural cavities are critical to balancing stability and activity of a membrane-integral enzyme. *Proc Natl Acad Sci U S A* **2020**, *117* (36), 22146-22156. DOI: 10.1073/pnas.1917770117.
- (94) Laskowski, R. A. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* **1995**, *13* (5), 323-330, 307-328. DOI: 10.1016/0263-7855(95)00073-9.
- (95) Berezovsky, I. N.; Shakhnovich, E. I. Physics and evolution of thermophilic adaptation. *Proc Natl Acad Sci U S A* **2005**, *102* (36), 12742-12747. DOI: 10.1073/pnas.0503890102.
- (96) Guarnera, E.; Berezovsky, I. N. Allosteric sites: remote control in regulation of protein activity. *Curr Opin Struct Biol* **2016**, *37*, 1-8. DOI: 10.1016/j.sbi.2015.10.004.

- (97) Wodak, S. J.; Paci, E.; Dokholyan, N. V.; Berezovsky, I. N.; Horovitz, A.; Li, J.; Hilser, V. J.; Bahar, I.; Karanicolas, J.; Stock, G.; et al. Allostery in Its Many Disguises: From Theory to Applications. *Structure* **2019**, *27* (4), 566-578. DOI: 10.1016/j.str.2019.01.003.
- (98) Ragland, D. A.; Whitfield, T. W.; Lee, S. K.; Swanstrom, R.; Zeldovich, K. B.; Kurt-Yilmaz, N.; Schiffer, C. A. Elucidating the Interdependence of Drug Resistance from Combinations of Mutations. *J Chem Theory Comput* **2017**, *13* (11), 5671-5682. DOI: 10.1021/acs.jctc.7b00601.
- (99) Ghosh, A. K.; Yu, X.; Osswald, H. L.; Agniswamy, J.; Wang, Y. F.; Amano, M.; Weber, I. T.; Mitsuya, H. Structure-based design of potent HIV-1 protease inhibitors with modified P1-biphenyl ligands: synthesis, biological evaluation, and enzyme-inhibitor X-ray structural studies. *J Med Chem* **2015**, *58* (13), 5334-5343. DOI: 10.1021/acs.jmedchem.5b00676.
- (100) Mittal, S.; Bandaranayake, R. M.; King, N. M.; Prabu-Jeyabalan, M.; Nalam, M. N.; Nalivaika, E. A.; Yilmaz, N. K.; Schiffer, C. A. Structural and thermodynamic basis of amprenavir/darunavir and atazanavir resistance in HIV-1 protease with mutations at residue 50. *J Virol* **2013**, *87* (8), 4176-4184. DOI: 10.1128/JVI.03486-12.
- (101) Tokuriki, N.; Oldfield, C. J.; Uversky, V. N.; Berezovsky, I. N.; Tawfik, D. S. Do viral proteins possess unique biophysical features? *Trends Biochem Sci* **2009**, *34* (2), 53-59. DOI: 10.1016/j.tibs.2008.10.009.
- (102) Zheng, F.; Vermaas, J. V.; Zheng, J.; Wang, Y.; Tu, T.; Wang, X.; Xie, X.; Yao, B.; Beckham, G. T.; Luo, H. Activity and Thermostability of GH5 Endoglucanase Chimeras from Mesophilic and Thermophilic Parents. *Appl Environ Microbiol* **2019**, *85* (5). DOI: 10.1128/AEM.02079-18.
- (103) Soh, L. M. J.; Mak, W. S.; Lin, P. P.; Mi, L.; Chen, F. Y.; Damoiseaux, R.; Siegel, J. B.; Liao, J. C. Engineering a Thermostable Keto Acid Decarboxylase Using Directed Evolution and Computationally Directed Protein Design. *ACS Synth Biol* **2017**, *6* (4), 610-618. DOI: 10.1021/acssynbio.6b00240.
- (104) Aqvist, J.; Kazemi, M.; Isaksen, G. V.; Brandsdal, B. O. Entropy and Enzyme Catalysis. *Acc Chem Res* **2017**, *50* (2), 199-207. DOI: 10.1021/acs.accounts.6b00321.
- (105) Vangone, A.; Bonvin, A. PRODIGY: A Contact-based Predictor of Binding Affinity in Protein-protein Complexes. *Bio Protoc* **2017**, *7* (3), e2124. DOI: 10.21769/BioProtoc.2124.
- (106) Callender, R.; Dyer, R. B. The dynamical nature of enzymatic catalysis. *Acc Chem Res* **2015**, *48* (2), 407-413. DOI: 10.1021/ar5002928.
- (107) Gibbs, A. C. Elements and modulation of functional dynamics. *J Med Chem* **2014**, *57* (19), 7819-7837. DOI: 10.1021/jm500325k.
- (108) Bigman, L. S.; Levy, Y. Stability Effects of Protein Mutations: The Role of Long-Range Contacts. *J Phys Chem B* **2018**, *122* (49), 11450-11459. DOI: 10.1021/acs.jpcc.8b07379.
- (109) Stetz, G.; Verkhivker, G. M. Computational Analysis of Residue Interaction Networks and Coevolutionary Relationships in the Hsp70 Chaperones: A Community-Hopping Model of Allosteric Regulation and Communication. *PLoS Comput Biol* **2017**, *13* (1), e1005299. DOI: 10.1371/journal.pcbi.1005299.
- (110) Pires, D. E.; Ascher, D. B.; Blundell, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res* **2014**, *42* (Web Server issue), W314-319. DOI: 10.1093/nar/gku411.
- (111) Dehghanpoor, R.; Ricks, E.; Hursh, K.; Gunderson, S.; Farhoodi, R.; Haspel, N.; Hutchinson, B.; Jagodzinski, F. Predicting the Effect of Single and Multiple Mutations on Protein Structural Stability. *Molecules* **2018**, *23* (2). DOI: 10.3390/molecules23020251.
- (112) Iannuzzelli, J. A.; Bacik, J. P.; Moore, E. J.; Shen, Z.; Irving, E. M.; Vargas, D. A.; Khare, S. D.; Ando, N.; Fasan, R. Tuning Enzyme Thermostability via Computationally Guided Covalent Stapling and Structural Basis of Enhanced Stabilization. *Biochemistry* **2022**, *61* (11), 1041-1054. DOI: 10.1021/acs.biochem.2c00033.
- (113) McAuley, M.; Timson, D. J. Modulating Mobility: a Paradigm for Protein Engineering? *Appl Biochem Biotechnol* **2017**, *181* (1), 83-90. DOI: 10.1007/s12010-016-2200-y.
- (114) Singh, M. K.; Streu, K.; McCrone, A. J.; Dominy, B. N. The evolution of catalytic function in the HIV-1 protease. *J Mol Biol* **2011**, *408* (4), 792-805. DOI: 10.1016/j.jmb.2011.02.031.
- (115) Tajiellyato, N.; Li, L.; Peng, Y.; Alper, J.; Alexov, E. E-hooks provide guidance and a soft landing for the microtubule binding domain of dynein. *Sci Rep* **2018**, *8* (1), 13266. DOI: 10.1038/s41598-018-31480-9.

- (116) Ung, P. M.; Ghanakota, P.; Graham, S. E.; Lexa, K. W.; Carlson, H. A. Identifying binding hot spots on protein surfaces by mixed-solvent molecular dynamics: HIV-1 protease as a test case. *Biopolymers* **2016**, *105* (1), 21-34. DOI: 10.1002/bip.22742.
- (117) Borges, B.; Gallo, G.; Coelho, C.; Negri, N.; Maiello, F.; Hardy, L.; Wurtele, M. Dynamic cross correlation analysis of *Thermus thermophilus* alkaline phosphatase and determinants of thermostability. *Biochim Biophys Acta Gen Subj* **2021**, *1865* (7), 129895. DOI: 10.1016/j.bbagen.2021.129895.
- (118) Kasahara, K.; Mohan, N.; Fukuda, I.; Nakamura, H. mDCC_tools: characterizing multi-modal atomic motions in molecular dynamics trajectories. *Bioinformatics* **2016**, *32* (16), 2531-2533. DOI: 10.1093/bioinformatics/btw129.
- (119) Kasahara, K.; Fukuda, I.; Nakamura, H. A novel approach of dynamic cross correlation analysis on molecular dynamics simulations and its application to Ets1 dimer-DNA complex. *PLoS One* **2014**, *9* (11), e112419. DOI: 10.1371/journal.pone.0112419.
- (120) Grant, B. J.; Rodrigues, A. P.; ElSawy, K. M.; McCammon, J. A.; Caves, L. S. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* **2006**, *22* (21), 2695-2696. DOI: 10.1093/bioinformatics/btl461.
- (121) Matko, J.; Tron, L.; Balazs, M.; Hevessy, J.; Somogyi, B.; Damjanovich, S. Correlation between activity and dynamics of the protein matrix of phosphorylase b. *Biochemistry* **1980**, *19* (25), 5782-5786. DOI: 10.1021/bi00566a018.
- (122) Hunenberger, P. H.; Mark, A. E.; van Gunsteren, W. F. Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. *J Mol Biol* **1995**, *252* (4), 492-503. DOI: 10.1006/jmbi.1995.0514.
- (123) Kaledin, M.; Brown, A.; Kaledin, A. L.; Bowman, J. M. Normal mode analysis using the driven molecular dynamics method. II. An application to biological macromolecules. *J Chem Phys* **2004**, *121* (12), 5646-5653. DOI: 10.1063/1.1777573.
- (124) Nagarajan, S.; Sakkiah, S. Exploring a potential allosteric inhibition mechanism in the motor domain of human Eg-5. *J Biomol Struct Dyn* **2019**, *37* (9), 2394-2403. DOI: 10.1080/07391102.2018.1486229.
- (125) Sogunmez, N.; Akten, E. D. Intrinsic Dynamics and Causality in Correlated Motions Unraveled in Two Distinct Inactive States of Human beta2-Adrenergic Receptor. *J Phys Chem B* **2019**, *123* (17), 3630-3642. DOI: 10.1021/acs.jpcc.9b01202.
- (126) Klinman, J. P.; Kohen, A. Evolutionary aspects of enzyme dynamics. *J Biol Chem* **2014**, *289* (44), 30205-30212. DOI: 10.1074/jbc.R114.565515.
- (127) Kornev, A. P.; Taylor, S. S. Dynamics-Driven Allostery in Protein Kinases. *Trends Biochem Sci* **2015**, *40* (11), 628-647. DOI: 10.1016/j.tibs.2015.09.002.
- (128) Radkiewicz, J. L.; Brooks, C. L. Protein Dynamics in Enzymatic Catalysis: Exploration of Dihydrofolate Reductase. *Journal of the American Chemical Society* **2000**, *122* (2), 225-231. DOI: 10.1021/ja9913838.
- (129) Rod, T. H.; Radkiewicz, J. L.; Brooks, C. L., 3rd. Correlated motion and the effect of distal mutations in dihydrofolate reductase. *Proc Natl Acad Sci U S A* **2003**, *100* (12), 6980-6985. DOI: 10.1073/pnas.1230801100.
- (130) Tofoleanu, F.; Yuan, Y.; Pickard, F. C. t.; Tywoniuk, B.; Brooks, B. R.; Buchete, N. V. Structural Modulation of Human Amylin Protofilaments by Naturally Occurring Mutations. *J Phys Chem B* **2018**, *122* (21), 5657-5665. DOI: 10.1021/acs.jpcc.7b12083.
- (131) Shannon, C. E. Communication theory of secrecy systems. 1945. *MD Comput* **1998**, *15* (1), 57-64.
- (132) Shannon, C. E. The mathematical theory of communication. 1963. *MD Comput* **1997**, *14* (4), 306-317.
- (133) Carley, D. W.; Shannon, D. C. Relative stability of human respiration during progressive hypoxia. *J Appl Physiol* (1985) **1988**, *65* (3), 1389-1399. DOI: 10.1152/jappl.1988.65.3.1389.
- (134) Aguilar, D.; Oliva, B.; Marino Buslje, C. Mapping the mutual information network of enzymatic families in the protein structure to unveil functional features. *PLoS One* **2012**, *7* (7), e41430. DOI: 10.1371/journal.pone.0041430.
- (135) Little, D. Y.; Chen, L. Identification of coevolving residues and coevolution potentials emphasizing structure, bond formation and catalytic coordination in protein evolution. *PLoS One* **2009**, *4* (3), e4762. DOI: 10.1371/journal.pone.0004762.

- (136) Shackelford, G.; Karplus, K. Contact prediction using mutual information and neural nets. *Proteins* **2007**, *69 Suppl 8*, 159-164. DOI: 10.1002/prot.21791.
- (137) McClendon, C. L.; Kornev, A. P.; Gilson, M. K.; Taylor, S. S. Dynamic architecture of a protein kinase. *Proc Natl Acad Sci U S A* **2014**, *111* (43), E4623-4631. DOI: 10.1073/pnas.1418402111.
- (138) Brandman, R.; Brandman, Y.; Pande, V. S. Sequence Coevolution between RNA and Protein Characterized by Mutual Information between Residue Triplets. *Plos One* **2012**, *7* (1). DOI: ARTN e30022 10.1371/journal.pone.0030022.
- (139) Marks, D. S.; Colwell, L. J.; Sheridan, R.; Hopf, T. A.; Pagnani, A.; Zecchina, R.; Sander, C. Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **2011**, *6* (12), e28766. DOI: 10.1371/journal.pone.0028766.
- (140) Weigt, M.; White, R. A.; Szurmant, H.; Hoch, J. A.; Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A* **2009**, *106* (1), 67-72. DOI: 10.1073/pnas.0805923106.
- (141) Ahlstrom, L. S.; Vorontsov, II; Shi, J.; Miyashita, O. Effect of the Crystal Environment on Side-Chain Conformational Dynamics in Cyanovirin-N Investigated through Crystal and Solution Molecular Dynamics Simulations. *PLoS One* **2017**, *12* (1), e0170337. DOI: 10.1371/journal.pone.0170337.
- (142) Frye, K. J.; Royer, C. A. Probing the contribution of internal cavities to the volume change of protein unfolding under pressure. *Protein Sci* **1998**, *7* (10), 2217-2222. DOI: 10.1002/pro.5560071020.
- (143) Isom, D. G.; Castaneda, C. A.; Cannon, B. R.; Garcia-Moreno, B. Large shifts in pKa values of lysine residues buried inside a protein. *Proc Natl Acad Sci U S A* **2011**, *108* (13), 5260-5265. DOI: 10.1073/pnas.1010750108.
- (144) Miller, C.; Davlieva, M.; Wilson, C.; White, K. I.; Counago, R.; Wu, G.; Myers, J. C.; Wittung-Stafshede, P.; Shamoo, Y. Experimental evolution of adenylate kinase reveals contrasting strategies toward protein thermostability. *Biophys J* **2010**, *99* (3), 887-896. DOI: 10.1016/j.bpj.2010.04.076.
- (145) Mohamed, E. I.; de Lorenzo, A. Modeling combined transport of water and test macromolecules across the glomerular capillary barrier: dynamics of the permselectivity. *Eur Biophys J* **2002**, *31* (3), 163-171. DOI: 10.1007/s00249-001-0203-1.
- (146) Csermely, P.; Korcsmaros, T.; Kiss, H. J.; London, G.; Nussinov, R. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol Ther* **2013**, *138* (3), 333-408. DOI: 10.1016/j.pharmthera.2013.01.016.
- (147) Vanangamudi, M.; Nair, P. C.; Engels, S. E. M.; Palaniappan, S.; Namasivayam, V. Structural Insights to Human Immunodeficiency Virus (HIV-1) Targets and Their Inhibition. *Adv Exp Med Biol* **2021**, *1322*, 63-95. DOI: 10.1007/978-981-16-0267-2_3.
- (148) Kleinpeter, A. B.; Freed, E. O. HIV-1 Maturation: Lessons Learned from Inhibitors. *Viruses* **2020**, *12* (9). DOI: 10.3390/v12090940.
- (149) Yu, F. H.; Wang, C. T. HIV-1 protease with leucine zipper fused at N-terminus exhibits enhanced linker amino acid-dependent activity. *Retrovirology* **2018**, *15* (1), 32. DOI: 10.1186/s12977-018-0413-6.
- (150) Sadeghi, L.; Lolaie, M.; Tabatabai, R. A.; Bayanolhagh, S.; Taj, L.; Ahmadi, N. E.; Abedinzadeh, N.; Vakili, F.; Ahmadi, G.; Mohraz, M. HIV-1 Drug Resistance Profiles for the HIV Protease and Reverse Transcriptase Gene in Patients Receiving Combination Therapy in Tehran, Iran. *Infect Disord Drug Targets* **2018**, *18* (3), 241-248. DOI: 10.2174/1871526518666180416110259.
- (151) Hosseini, A.; Alibes, A.; Noguera-Julian, M.; Gil, V.; Paredes, R.; Soliva, R.; Orozco, M.; Guallar, V. Computational Prediction of HIV-1 Resistance to Protease Inhibitors. *J Chem Inf Model* **2016**, *56* (5), 915-923. DOI: 10.1021/acs.jcim.5b00667.
- (152) Gashnikova, N. M.; Astakhova, E. M.; Gashnikova, M. P.; Bocharov, E. F.; Petrova, S. V.; Pun'ko, O. A.; Popkov, A. V.; Totmenin, A. V. HIV-1 Epidemiology, Genetic Diversity, and Primary Drug Resistance in the Tyumen Oblast, Russia. *Biomed Res Int* **2016**, *2016*, 2496280. DOI: 10.1155/2016/2496280.
- (153) Wensing, A. M.; Calvez, V.; Gunthard, H. F.; Johnson, V. A.; Paredes, R.; Pillay, D.; Shafer, R. W.; Richman, D. D. 2015 Update of the Drug Resistance Mutations in HIV-1. *Top Antivir Med* **2015**, *23* (4), 132-141.
- (154) Sangeda, R. Z.; Theys, K.; Beheydt, G.; Rhee, S. Y.; Deforche, K.; Vercauteren, J.; Libin, P.; Imbrechts, S.; Grossman, Z.; Camacho, R. J.; et al. HIV-1 fitness landscape models for indinavir treatment

pressure using observed evolution in longitudinal sequence data are predictive for treatment failure. *Infect Genet Evol* **2013**, *19*, 349-360. DOI: 10.1016/j.meegid.2013.03.014.

(155) Wah Tan, Z.; Tee, W. V.; Berezovsky, I. N. Learning About Allosteric Drugs and Ways to Design Them. *J Mol Biol* **2022**, *434* (17), 167692. DOI: 10.1016/j.jmb.2022.167692.

(156) Boucher, J. I.; Whitfield, T. W.; Dauphin, A.; Nachum, G.; Hollins, C.; Zeldovich, K. B.; Swanstrom, R.; Schiffer, C. A.; Luban, J.; Bolon, D. N. A. Constrained Mutational Sampling of Amino Acids in HIV-1 Protease Evolution. *Mol Biol Evol* **2019**, *36* (4), 798-810. DOI: 10.1093/molbev/msz022.

(157) Ahuja, L. G.; Taylor, S. S.; Kornev, A. P. Tuning the "violin" of protein kinases: The role of dynamics-based allostery. *IUBMB Life* **2019**, *71* (6), 685-696. DOI: 10.1002/iub.2057.

(158) Stock, G.; Hamm, P. A non-equilibrium approach to allosteric communication. *Philos Trans R Soc Lond B Biol Sci* **2018**, *373* (1749). DOI: 10.1098/rstb.2017.0187.

(159) Saavedra, H. G.; Wrabl, J. O.; Anderson, J. A.; Li, J.; Hilser, V. J. Dynamic allostery can drive cold adaptation in enzymes. *Nature* **2018**, *558* (7709), 324-328. DOI: 10.1038/s41586-018-0183-2.

(160) Otten, R.; Liu, L.; Kenner, L. R.; Clarkson, M. W.; Mavor, D.; Tawfik, D. S.; Kern, D.; Fraser, J. S. Rescue of conformational dynamics in enzyme catalysis by directed evolution. *Nat Commun* **2018**, *9* (1), 1314. DOI: 10.1038/s41467-018-03562-9.

(161) Kurkcuoglu, O. Exploring allosteric communication in multiple states of the bacterial ribosome using residue network analysis. *Turk J Biol* **2018**, *42* (5), 392-404. DOI: 10.3906/biy-1802-77.

(162) Kim, T. H.; Mehrabi, P.; Ren, Z.; Sljoka, A.; Ing, C.; Bezginov, A.; Ye, L.; Pomes, R.; Prosser, R. S.; Pai, E. F. The role of dimer asymmetry and protomer dynamics in enzyme catalysis. *Science* **2017**, *355* (6322). DOI: 10.1126/science.aag2355.

(163) Jimenez-Oses, G.; Osuna, S.; Gao, X.; Sawaya, M. R.; Gilson, L.; Collier, S. J.; Huisman, G. W.; Yeates, T. O.; Tang, Y.; Houk, K. N. The role of distant mutations and allosteric regulation on LovD active site dynamics. *Nat Chem Biol* **2014**, *10* (6), 431-436. DOI: 10.1038/nchembio.1503.

(164) Lange, O. F.; Grubmüller, H. Full correlation analysis of conformational protein dynamics. *Proteins* **2008**, *70* (4), 1294-1312. DOI: 10.1002/prot.21618.

(165) Bradley, M. J.; Chivers, P. T.; Baker, N. A. Molecular dynamics simulation of the Escherichia coli NikR protein: equilibrium conformational fluctuations reveal interdomain allosteric communication pathways. *J Mol Biol* **2008**, *378* (5), 1155-1173. DOI: 10.1016/j.jmb.2008.03.010.

(166) Ahuja, L. G.; Kornev, A. P.; McClendon, C. L.; Veglia, G.; Taylor, S. S. Mutation of a kinase allosteric node uncouples dynamics linked to phosphotransfer. *Proc Natl Acad Sci U S A* **2017**, *114* (6), E931-E940. DOI: 10.1073/pnas.1620667114.

(167) Abrusan, G.; Marsh, J. A. Ligand-Binding-Site Structure Shapes Allosteric Signal Transduction and the Evolution of Allostery in Protein Complexes. *Mol Biol Evol* **2019**, *36* (8), 1711-1727. DOI: 10.1093/molbev/msz093.

(168) Buchenberg, S.; Sittel, F.; Stock, G. Time-resolved observation of protein allosteric communication. *Proc Natl Acad Sci U S A* **2017**, *114* (33), E6804-E6811. DOI: 10.1073/pnas.1707694114.

(169) Tang, S.; Liao, J. C.; Dunn, A. R.; Altman, R. B.; Spudich, J. A.; Schmidt, J. P. Predicting allosteric communication in myosin via a pathway of conserved residues. *J Mol Biol* **2007**, *373* (5), 1361-1373. DOI: 10.1016/j.jmb.2007.08.059.

(170) Nucci, N. V.; Fuglestad, B.; Athanasoula, E. A.; Wand, A. J. Role of cavities and hydration in the pressure unfolding of T4 lysozyme. *Proc Natl Acad Sci U S A* **2014**, *111* (38), 13846-13851. DOI: 10.1073/pnas.1410655111.

(171) Roche, J.; Dellarole, M.; Caro, J. A.; Norberto, D. R.; Garcia, A. E.; Garcia-Moreno, B.; Roumestand, C.; Royer, C. A. Effect of internal cavities on folding rates and routes revealed by real-time pressure-jump NMR spectroscopy. *J Am Chem Soc* **2013**, *135* (39), 14610-14618. DOI: 10.1021/ja406682e.

(172) Dellarole, M.; Kobayashi, K.; Rouget, J. B.; Caro, J. A.; Roche, J.; Islam, M. M.; Garcia-Moreno, E. B.; Kuroda, Y.; Royer, C. A. Probing the physical determinants of thermal expansion of folded proteins. *J Phys Chem B* **2013**, *117* (42), 12742-12749. DOI: 10.1021/jp401113p.

(173) van Domselaar, R.; Njenda, D. T.; Rao, R.; Sonnerborg, A.; Singh, K.; Neogi, U. HIV-1 Subtype C with PYx_E Insertion Has Enhanced Binding of Gag-p6 to Host Cell Protein ALIX and Increased Replication Fitness. *J Virol* **2019**, *93* (9). DOI: 10.1128/JVI.00077-19.

- (174) Koss, H.; Bunney, T. D.; Esposito, D.; Martins, M.; Katan, M.; Driscoll, P. C. Dynamic Allostery in PLCgamma1 and Its Modulation by a Cancer Mutation Revealed by MD Simulation and NMR. *Biophys J* **2018**, *115* (1), 31-45. DOI: 10.1016/j.bpj.2018.05.031.
- (175) Keighobadi, M.; Emami, S.; Lagzian, M.; Fakhari, M.; Rafiei, A.; Valadan, R. Molecular Modeling and Structural Stability of Wild-Type and Mutant CYP51 from *Leishmania major*: In Vitro and In Silico Analysis of a Laboratory Strain. *Molecules* **2018**, *23* (3). DOI: 10.3390/molecules23030696.
- (176) Garbelli, A.; Riva, V.; Crespan, E.; Maga, G. How to win the HIV-1 drug resistance hurdle race: running faster or jumping higher? *Biochem J* **2017**, *474* (10), 1559-1577. DOI: 10.1042/BCJ20160772.
- (177) Namanja, A. T.; Wang, J.; Buettner, R.; Colson, L.; Chen, Y. Allosteric Communication across STAT3 Domains Associated with STAT3 Function and Disease-Causing Mutation. *J Mol Biol* **2016**, *428* (3), 579-589. DOI: 10.1016/j.jmb.2016.01.003.
- (178) George, D. C.; Chakraborty, C.; Haneef, S. A.; Nagasundaram, N.; Chen, L.; Zhu, H. Evolution- and structure-based computational strategy reveals the impact of deleterious missense mutations on MODY 2 (maturity-onset diabetes of the young, type 2). *Theranostics* **2014**, *4* (4), 366-385. DOI: 10.7150/thno.7473.
- (179) Matsukuma, S.; Yoshihara, M.; Kasai, F.; Kato, A.; Yoshida, A.; Akaike, M.; Kobayashi, O.; Nakayama, H.; Sakuma, Y.; Yoshida, T.; et al. Rapid and simple detection of hot spot point mutations of epidermal growth factor receptor, BRAF, and NRAS in cancers using the loop-hybrid mobility shift assay. *J Mol Diagn* **2006**, *8* (4), 504-512. DOI: 10.2353/jmoldx.2006.060030.
- (180) Pachetti, M.; Marini, B.; Benedetti, F.; Giudici, F.; Mauro, E.; Storici, P.; Masciovecchio, C.; Angeletti, S.; Ciccozzi, M.; Gallo, R. C.; et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J Transl Med* **2020**, *18* (1), 179. DOI: 10.1186/s12967-020-02344-6.
- (181) Pfaller, M. A. Antifungal drug resistance: mechanisms, epidemiology, and consequences for treatment. *Am J Med* **2012**, *125* (1 Suppl), S3-13. DOI: 10.1016/j.amjmed.2011.11.001.
- (182) Konnyu, B.; Sadiq, S. K.; Turanyi, T.; Hirmondo, R.; Muller, B.; Krausslich, H. G.; Coveney, P. V.; Muller, V. Gag-Pol processing during HIV-1 virion maturation: a systems biology approach. *PLoS Comput Biol* **2013**, *9* (6), e1003103. DOI: 10.1371/journal.pcbi.1003103.
- (183) Delatorre, E.; Bello, G. Phylodynamics of the HIV-1 epidemic in Cuba. *PLoS One* **2013**, *8* (9), e72448. DOI: 10.1371/journal.pone.0072448.
- (184) Kousiappa, I.; Achilleos, C.; Hezka, J.; Lazarou, Y.; Othonos, K.; Demetriades, I.; Kostrikis, L. G. Molecular characterization of HIV type 1 strains from newly diagnosed patients in Cyprus (2007-2009) recovers multiple clades including unique recombinant strains and lack of transmitted drug resistance. *AIDS Res Hum Retroviruses* **2011**, *27* (11), 1183-1199. DOI: 10.1089/AID.2011.0060.
- (185) Ozer, N.; Ozen, A.; Schiffer, C. A.; Haliloglu, T. Drug-resistant HIV-1 protease regains functional dynamics through cleavage site coevolution. *Evol Appl* **2015**, *8* (2), 185-198. DOI: 10.1111/eva.12241.
- (186) Goodsell, D. S.; Olson, A. J. Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct* **2000**, *29*, 105-153. DOI: 10.1146/annurev.biophys.29.1.105.
- (187) Jackel, C.; Kast, P.; Hilvert, D. Protein design by directed evolution. *Annu Rev Biophys* **2008**, *37*, 153-173. DOI: 10.1146/annurev.biophys.37.032807.125832.
- (188) Huang, J.; Koide, A.; Makabe, K.; Koide, S. Design of protein function leaps by directed domain interface evolution. *Proc Natl Acad Sci U S A* **2008**, *105* (18), 6578-6583. DOI: 10.1073/pnas.0801097105.
- (189) Levy, E. D.; Boeri Erba, E.; Robinson, C. V.; Teichmann, S. A. Assembly reflects evolution of protein complexes. *Nature* **2008**, *453* (7199), 1262-1265. DOI: 10.1038/nature06942.
- (190) Bahadur, R. P.; Chakrabarti, P.; Rodier, F.; Janin, J. Dissecting subunit interfaces in homodimeric proteins. *Proteins* **2003**, *53* (3), 708-719. DOI: 10.1002/prot.10461.
- (191) Shafer, R. W. Human Immunodeficiency Virus Type 1 Drug Resistance Mutations Update. *J Infect Dis* **2017**, *216* (suppl_9), S843-S846. DOI: 10.1093/infdis/jix398.
- (192) Johnson, V. A.; Calvez, V.; Gunthard, H. F.; Paredes, R.; Pillay, D.; Shafer, R.; Wensing, A. M.; Richman, D. D. 2011 update of the drug resistance mutations in HIV-1. *Top Antivir Med* **2011**, *19* (4), 156-164.
- (193) Johnson, V. A.; Brun-Vezinet, F.; Clotet, B.; Gunthard, H. F.; Kuritzkes, D. R.; Pillay, D.; Schapiro, J. M.; Richman, D. D. Update of the drug resistance mutations in HIV-1: December 2010. *Top HIV Med* **2010**, *18* (5), 156-163.

- (194) Ichiye, T.; Karplus, M. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins* **1991**, *11* (3), 205-217. DOI: 10.1002/prot.340110305.
- (195) Henzler-Wildman, K. A.; Thai, V.; Lei, M.; Ott, M.; Wolf-Watz, M.; Fenn, T.; Pozharski, E.; Wilson, M. A.; Petsko, G. A.; Karplus, M.; et al. Intrinsic motions along an enzymatic reaction trajectory. *Nature* **2007**, *450* (7171), 838-844. DOI: 10.1038/nature06410.
- (196) Kitao, A.; Go, N. Investigating protein dynamics in collective coordinate space. *Curr Opin Struct Biol* **1999**, *9* (2), 164-169. DOI: 10.1016/S0959-440X(99)80023-2.
- (197) Balsera, M. A.; Wriggers, W.; Oono, Y.; Schulten, K. Principal Component Analysis and Long Time Protein Dynamics. *The Journal of Physical Chemistry* **1996**, *100* (7), 2567-2572. DOI: 10.1021/jp9536920.
- (198) Karplus, M.; Ichiye, T. Comment on a "fluctuation and cross correlation analysis of protein motions observed in nanosecond molecular dynamics simulations". *J Mol Biol* **1996**, *263* (2), 120-122. DOI: 10.1006/jmbi.1996.0562.
- (199) Jilani, M.; Turcan, A.; Haspel, N.; Jagodzinski, F. Elucidating the Structural Impacts of Protein InDels. *Biomolecules* **2022**, *12* (10). DOI: 10.3390/biom12101435.
- (200) Zhao, Q. Molecular and thermodynamic mechanisms for protein adaptation. *Eur Biophys J* **2022**, *51* (7-8), 519-534. DOI: 10.1007/s00249-022-01618-9.
- (201) Gurung, A. B.; Bhattacharjee, A. Impact of a non-synonymous Q281R polymorphism on structure of human Lipoprotein-Associated Phospholipase A2 (Lp-PLA2). *J Cell Biochem* **2018**, *119* (8), 7009-7021. DOI: 10.1002/jcb.26909.
- (202) Appadurai, R.; Senapati, S. How Mutations Can Resist Drug Binding yet Keep HIV-1 Protease Functional. *Biochemistry* **2017**, *56* (23), 2907-2920. DOI: 10.1021/acs.biochem.7b00139.
- (203) Biedermannova, L.; Prokop, Z.; Gora, A.; Chovancova, E.; Kovacs, M.; Damborsky, J.; Wade, R. C. A single mutation in a tunnel to the active site changes the mechanism and kinetics of product release in haloalkane dehalogenase LinB. *J Biol Chem* **2012**, *287* (34), 29062-29074. DOI: 10.1074/jbc.M112.377853.
- (204) Ghosh, A. K.; Rao, K. V.; Nyalapatla, P. R.; Osswald, H. L.; Martyr, C. D.; Aoki, M.; Hayashi, H.; Agniswamy, J.; Wang, Y. F.; Bulut, H.; et al. Design and Development of Highly Potent HIV-1 Protease Inhibitors with a Crown-Like Oxotricyclic Core as the P2-Ligand To Combat Multidrug-Resistant HIV Variants. *J Med Chem* **2017**, *60* (10), 4267-4278. DOI: 10.1021/acs.jmedchem.7b00172.
- (205) Szekrenyi, A.; Garrahou, X.; Parella, T.; Joglar, J.; Bujons, J.; Clapes, P. Asymmetric assembly of aldose carbohydrates from formaldehyde and glyceraldehyde by tandem biocatalytic aldol reactions. *Nat Chem* **2015**, *7* (9), 724-729. DOI: 10.1038/nchem.2321.
- (206) Lucas, X.; Gunther, S. Using chiral molecules as an approach to address low-druggability recognition sites. *J Comput Chem* **2014**, *35* (29), 2114-2121. DOI: 10.1002/jcc.23726.
- (207) Pietropaolo, A.; Branduardi, D.; Bonomi, M.; Parrinello, M. A chirality-based metrics for free-energy calculations in biomolecular systems. *J Comput Chem* **2011**, *32* (12), 2627-2637. DOI: 10.1002/jcc.21842.
- (208) Prabu-Jeyabalan, M.; Nalivaika, E.; Schiffer, C. A. Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes. *Structure* **2002**, *10* (3), 369-381.
- (209) Kempf, D. J.; Marsh, K. C.; Denissen, J. F.; McDonald, E.; Vasavanonda, S.; Flentge, C. A.; Green, B. E.; Fino, L.; Park, C. H.; Kong, X. P.; et al. ABT-538 is a potent inhibitor of human immunodeficiency virus protease and has high oral bioavailability in humans. *Proc Natl Acad Sci U S A* **1995**, *92* (7), 2484-2488. DOI: 10.1073/pnas.92.7.2484.
- (210) Rhee, S. Y.; Gonzales, M. J.; Kantor, R.; Betts, B. J.; Ravela, J.; Shafer, R. W. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res* **2003**, *31* (1), 298-303. DOI: 10.1093/nar/gkg100.
- (211) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79* (2), 926-935. DOI: 10.1063/1.445869.
- (212) Singh, M. K.; Dominy, B. N. The evolution of cefotaximase activity in the TEM beta-lactamase. *J Mol Biol* **2012**, *415* (1), 205-220. DOI: 10.1016/j.jmb.2011.10.041.

- (213) Brooks, B. R.; Brooks, C. L., 3rd; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. CHARMM: the biomolecular simulation program. *J Comput Chem* **2009**, *30* (10), 1545-1614. DOI: 10.1002/jcc.21287.
- (214) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* **1998**, *102* (18), 3586-3616. DOI: 10.1021/jp973084f.
- (215) van Gunsteren, W. F.; Berendsen, H. J. C. Algorithms for macromolecular dynamics and constraint dynamics. *Molecular Physics* **1977**, *34* (5), 1311-1327. DOI: 10.1080/00268977700102571.
- (216) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. Constant pressure molecular dynamics simulation: The Langevin piston method. *The Journal of Chemical Physics* **1995**, *103* (11), 4613-4621. DOI: 10.1063/1.470648.
- (217) Benz, R. W.; Castro-Roman, F.; Tobias, D. J.; White, S. H. Experimental validation of molecular dynamics simulations of lipid bilayers: a new approach. *Biophys J* **2005**, *88* (2), 805-817. DOI: 10.1529/biophysj.104.046821.
- (218) Lee, J.; Cheng, X.; Swails, J. M.; Yeom, M. S.; Eastman, P. K.; Lemkul, J. A.; Wei, S.; Buckner, J.; Jeong, J. C.; Qi, Y.; et al. CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J Chem Theory Comput* **2016**, *12* (1), 405-413. DOI: 10.1021/acs.jctc.5b00935.
- (219) Pahari, S.; Sun, L.; Basu, S.; Alexov, E. DelPhiPKa: Including salt in the calculations and enabling polar residues to titrate. *Proteins* **2018**, *86* (12), 1277-1283. DOI: 10.1002/prot.25608.
- (220) Peng, Y.; Alexov, E. Computational investigation of proton transfer, pKa shifts and pH-optimum of protein-DNA and protein-RNA complexes. *Proteins* **2017**, *85* (2), 282-295. DOI: 10.1002/prot.25221.
- (221) Wang, L.; Zhang, M.; Alexov, E. DelPhiPKa web server: predicting pKa of proteins, RNAs and DNAs. *Bioinformatics* **2016**, *32* (4), 614-615. DOI: 10.1093/bioinformatics/btv607.
- (222) Hsin, J.; Arkhipov, A.; Yin, Y.; Stone, J. E.; Schulten, K. Using VMD: an introductory tutorial. *Curr Protoc Bioinformatics* **2008**, Chapter 5, Unit 5 7. DOI: 10.1002/0471250953.bi0507s24.
- (223) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J Mol Graph* **1996**, *14* (1), 33-38, 27-38. DOI: 10.1016/0263-7855(96)00018-5.
- (224) Verkhivker, G. M.; Agajanian, S.; Hu, G.; Tao, P. Allosteric Regulation at the Crossroads of New Technologies: Multiscale Modeling, Networks, and Machine Learning. *Front Mol Biosci* **2020**, *7*, 136. DOI: 10.3389/fmolb.2020.00136.
- (225) Chen, K.; Duan, W.; Han, Q.; Sun, X.; Li, W.; Hu, S.; Wan, J.; Wu, J.; Ge, Y.; Liu, D. Identification of the hot spot residues for pyridine derivative inhibitor CCT251455 and ATP substrate binding on monopolar spindle 1 (MPS1) kinase by molecular dynamic simulation. *J Biomol Struct Dyn* **2019**, *37* (3), 611-622. DOI: 10.1080/07391102.2018.1433552.
- (226) Miotto, M.; Di Rienzo, L.; Gosti, G.; Bo, L.; Parisi, G.; Piacentini, R.; Boffi, A.; Ruocco, G.; Milanetti, E. Inferring the stabilization effects of SARS-CoV-2 variants on the binding with ACE2 receptor. *Commun Biol* **2022**, *5* (1), 20221. DOI: 10.1038/s42003-021-02946-w.
- (227) Fariselli, P.; Martelli, P. L.; Savojardo, C.; Casadio, R. INPS: predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinformatics* **2015**, *31* (17), 2816-2821. DOI: 10.1093/bioinformatics/btv291.
- (228) Jimenez-Garcia, B.; Elez, K.; Koukos, P. I.; Bonvin, A. M.; Vangone, A. PRODIGY-crystal: a web-tool for classification of biological interfaces in protein complexes. *Bioinformatics* **2019**, *35* (22), 4821-4823. DOI: 10.1093/bioinformatics/btz437.
- (229) Xue, L. C.; Rodrigues, J. P.; Kastiris, P. L.; Bonvin, A. M.; Vangone, A. PRODIGY: a web server for predicting the binding affinity of protein-protein complexes. *Bioinformatics* **2016**, *32* (23), 3676-3678. DOI: 10.1093/bioinformatics/btw514.
- (230) Trachtman, H.; Nelson, P.; Adler, S.; Campbell, K. N.; Chaudhuri, A.; Derebail, V. K.; Gambaro, G.; Gesualdo, L.; Gipson, D. S.; Hogan, J.; et al. DUET: A Phase 2 Study Evaluating the Efficacy and Safety of Sparsentan in Patients with FSGS. *J Am Soc Nephrol* **2018**, *29* (11), 2745-2754. DOI: 10.1681/ASN.2018010091.

- (231) Montanucci, L.; Capriotti, E.; Birolo, G.; Benevenuta, S.; Pancotti, C.; Lal, D.; Fariselli, P. DDGun: an untrained predictor of protein stability changes upon amino acid variants. *Nucleic Acids Res* **2022**, *50* (W1), W222-227. DOI: 10.1093/nar/gkac325.
- (232) Montanucci, L.; Capriotti, E.; Frank, Y.; Ben-Tal, N.; Fariselli, P. DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. *BMC Bioinformatics* **2019**, *20* (Suppl 14), 335. DOI: 10.1186/s12859-019-2923-1.
- (233) Pires, D. E.; Ascher, D. B.; Blundell, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **2014**, *30* (3), 335-342. DOI: 10.1093/bioinformatics/btt691.
- (234) Pandurangan, A. P.; Ochoa-Montano, B.; Ascher, D. B.; Blundell, T. L. SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res* **2017**, *45* (W1), W229-W235. DOI: 10.1093/nar/gkx439.
- (235) Savojardo, C.; Fariselli, P.; Martelli, P. L.; Casadio, R. INPS-MD: a web server to predict stability of protein variants from sequence and structure. *Bioinformatics* **2016**, *32* (16), 2542-2544. DOI: 10.1093/bioinformatics/btw192.
- (236) Gonzalez-Andrade, M.; Rodriguez-Sotres, R.; Madariaga-Mazon, A.; Rivera-Chavez, J.; Mata, R.; Sosa-Peinado, A.; Del Pozo-Yauner, L.; Arias, O., II. Insights into molecular interactions between CaM and its inhibitors from molecular dynamics simulations and experimental data. *J Biomol Struct Dyn* **2016**, *34* (1), 78-91. DOI: 10.1080/07391102.2015.1022225.
- (237) Voit-Ostricki, L.; Lovas, S.; Watts, C. R. Conformation and Domain Movement Analysis of Human Matrix Metalloproteinase-2: Role of Associated Zn(2+) and Ca(2+) Ions. *Int J Mol Sci* **2019**, *20* (17). DOI: 10.3390/ijms20174194.
- (238) Wright, D. W.; Hall, B. A.; Kenway, O. A.; Jha, S.; Coveney, P. V. Computing Clinically Relevant Binding Free Energies of HIV-1 Protease Inhibitors. *J Chem Theory Comput* **2014**, *10* (3), 1228-1241. DOI: 10.1021/ct4007037.
- (239) Tokuriki, N.; Stricher, F.; Schymkowitz, J.; Serrano, L.; Tawfik, D. S. The stability effects of protein mutations appear to be universally distributed. *J Mol Biol* **2007**, *369* (5), 1318-1332. DOI: 10.1016/j.jmb.2007.03.069.
- (240) Bloom, J. D.; Labthavikul, S. T.; Otey, C. R.; Arnold, F. H. Protein stability promotes evolvability. *Proc Natl Acad Sci U S A* **2006**, *103* (15), 5869-5874. DOI: 10.1073/pnas.0510098103.
- (241) Agarwal, P. K.; Geist, A.; Gorin, A. Protein dynamics and enzymatic catalysis: investigating the peptidyl-prolyl cis-trans isomerization activity of cyclophilin A. *Biochemistry* **2004**, *43* (33), 10605-10618. DOI: 10.1021/bi0495228.
- (242) Eisenmesser, E. Z.; Millet, O.; Labeikovsky, W.; Korzhnev, D. M.; Wolf-Watz, M.; Bosco, D. A.; Skaliky, J. J.; Kay, L. E.; Kern, D. Intrinsic dynamics of an enzyme underlies catalysis. *Nature* **2005**, *438* (7064), 117-121. DOI: 10.1038/nature04105.
- (243) Hammes-Schiffer, S.; Benkovic, S. J. Relating protein motion to catalysis. *Annu Rev Biochem* **2006**, *75*, 519-541. DOI: 10.1146/annurev.biochem.75.103004.142800.
- (244) Klinman, J. P. Linking protein structure and dynamics to catalysis: the role of hydrogen tunnelling. *Philos Trans R Soc Lond B Biol Sci* **2006**, *361* (1472), 1323-1331. DOI: 10.1098/rstb.2006.1870.
- (245) Klinman, J. P. Enzyme dynamics: Control of active-site compression. *Nat Chem* **2010**, *2* (11), 907-909. DOI: 10.1038/nchem.886.
- (246) Boekelheide, N.; Salomon-Ferrer, R.; Miller, T. F., 3rd. Dynamics and dissipation in enzyme catalysis. *Proc Natl Acad Sci U S A* **2011**, *108* (39), 16159-16163. DOI: 10.1073/pnas.1106397108.
- (247) Hay, S.; Scrutton, N. S. Good vibrations in enzyme-catalysed reactions. *Nat Chem* **2012**, *4* (3), 161-168. DOI: 10.1038/nchem.1223.
- (248) Campbell, E.; Kaltenbach, M.; Correy, G. J.; Carr, P. D.; Porebski, B. T.; Livingstone, E. K.; Afriat-Jurnou, L.; Buckle, A. M.; Weik, M.; Hollfelder, F.; et al. The role of protein dynamics in the evolution of new enzyme function. *Nat Chem Biol* **2016**, *12* (11), 944-950. DOI: 10.1038/nchembio.2175.
- (249) Henzler-Wildman, K.; Kern, D. Dynamic personalities of proteins. *Nature* **2007**, *450* (7172), 964-972. DOI: 10.1038/nature06522.
- (250) Daniel, R. M.; Dunn, R. V.; Finney, J. L.; Smith, J. C. The role of dynamics in enzyme activity. *Annu Rev Biophys Biomol Struct* **2003**, *32*, 69-92. DOI: 10.1146/annurev.biophys.32.110601.142445.

- (251) Eisenmesser, E. Z.; Bosco, D. A.; Akke, M.; Kern, D. Enzyme dynamics during catalysis. *Science* **2002**, 295 (5559), 1520-1523. DOI: 10.1126/science.1066176.
- (252) Kern, D.; Eisenmesser, E. Z.; Wolf-Watz, M. Enzyme dynamics during catalysis measured by NMR spectroscopy. *Methods Enzymol* **2005**, 394, 507-524. DOI: 10.1016/S0076-6879(05)94021-4.
- (253) Wand, A. J. On the dynamic origins of allosteric activation. *Science* **2001**, 293 (5534), 1395. DOI: 10.1126/science.293.5534.1395a.
- (254) Cheng, Y.; Zhang, Y.; McCammon, J. A. How does activation loop phosphorylation modulate catalytic activity in the cAMP-dependent protein kinase: a theoretical study. *Protein Sci* **2006**, 15 (4), 672-683. DOI: 10.1110/ps.051852306.
- (255) Palmai, Z.; Chaloin, L.; Lionne, C.; Fidy, J.; Perahia, D.; Balog, E. Substrate binding modifies the hinge bending characteristics of human 3-phosphoglycerate kinase: a molecular dynamics study. *Proteins* **2009**, 77 (2), 319-329. DOI: 10.1002/prot.22437.
- (256) Schuyler, A. D.; Carlson, H. A.; Feldman, E. L. Computational methods for predicting sites of functionally important dynamics. *J Phys Chem B* **2009**, 113 (19), 6613-6622. DOI: 10.1021/jp808736c.
- (257) Watney, J. B.; Hammes-Schiffer, S. Comparison of coupled motions in Escherichia coli and Bacillus subtilis dihydrofolate reductase. *J Phys Chem B* **2006**, 110 (20), 10130-10138. DOI: 10.1021/jp0605956.
- (258) Ishida, T. Effects of point mutation on enzymatic activity: correlation between protein electronic structure and motion in chorismate mutase reaction. *J Am Chem Soc* **2010**, 132 (20), 7104-7118. DOI: 10.1021/ja100744h.
- (259) Sk, M. F.; Roy, R.; Kar, P. Exploring the potency of currently used drugs against HIV-1 protease of subtype D variant by using multiscale simulations. *J Biomol Struct Dyn* **2021**, 39 (3), 988-1003. DOI: 10.1080/07391102.2020.1724196.
- (260) Borisova, A. S.; Eneyskaya, E. V.; Jana, S.; Badino, S. F.; Kari, J.; Amore, A.; Karlsson, M.; Hansson, H.; Sandgren, M.; Himmel, M. E.; et al. Correlation of structure, function and protein dynamics in GH7 cellobiohydrolases from Trichoderma atroviride, T. reesei and T. harzianum. *Biotechnol Biofuels* **2018**, 11, 5. DOI: 10.1186/s13068-017-1006-7.
- (261) Fernandez, A.; Tawfik, D. S.; Berkhout, B.; Sanders, R.; Kloczkowski, A.; Sen, T.; Jernigan, B. Protein promiscuity: drug resistance and native functions--HIV-1 case. *J Biomol Struct Dyn* **2005**, 22 (6), 615-624. DOI: 10.1080/07391102.2005.10531228.
- (262) Olguin-Rodriguez, P. V.; Arzate-Mena, J. D.; Corsi-Cabrera, M.; Gast, H.; Marin-Garcia, A.; Mathis, J.; Ramos Loyo, J.; Del Rio-Portilla, I. Y.; Rummel, C.; Schindler, K.; et al. Characteristic Fluctuations Around Stable Attractor Dynamics Extracted from Highly Nonstationary Electroencephalographic Recordings. *Brain Connect* **2018**, 8 (8), 457-474. DOI: 10.1089/brain.2018.0609.
- (263) Labas, A.; Szabo, E.; Mones, L.; Fuxreiter, M. Optimization of reorganization energy drives evolution of the designed Kemp eliminase KE07. *Biochim Biophys Acta* **2013**, 1834 (5), 908-917. DOI: 10.1016/j.bbapap.2013.01.005.
- (264) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *Journal of Chemical Physics* **1983**, 79 (2), 926-935. DOI: Doi 10.1063/1.445869.
- (265) Feller, S. E.; Zhang, Y. H.; Pastor, R. W.; Brooks, B. R. Constant-Pressure Molecular-Dynamics Simulation - the Langevin Piston Method. *Journal of Chemical Physics* **1995**, 103 (11), 4613-4621. DOI: Doi 10.1063/1.470648.
- (266) Fontoura, C. A.; Castellani, G.; Mombach, J. C. The R implementation of the CRAN package PATHChange, a tool to study genetic pathway alterations in transcriptomic data. *Comput Biol Med* **2016**, 78, 76-80. DOI: 10.1016/j.combiomed.2016.09.010.
- (267) Hornik, K. The Comprehensive R Archive Network. *Wires Comput Stat* **2012**, 4 (4), 394-398. DOI: 10.1002/wics.1212.
- (268) Unlu, A.; Yanagida, T. R you ready for R?: the CRAN psychometrics task view. *Br J Math Stat Psychol* **2011**, 64 (Pt 1), 182-186. DOI: 10.1348/000711010X519320.
- (269) Grant, B. J.; Skjaerven, L.; Yao, X. Q. The Bio3D packages for structural bioinformatics. *Protein Sci* **2021**, 30 (1), 20-30. DOI: 10.1002/pro.3923.
- (270) Jerome Friedman, T. H., Robert Tibshirani Regularization Paths for

- Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1-22. . **2010**.
- (271) Siriseriwan, W. smotefamily: A Collection of Oversampling Techniques for Class Imbalance Problem Based on SMOTE. R package version 1.3.1. . **2019**.
- (272) *The Elements of Statistical Learning*; 2009. DOI: 10.1007/978-0-387-84858-7.
- (273) Sidhu, A.; Grosbart, M.; Sanchez, H.; Verhagen, B.; van der Zon, N. L. L.; Ristic, D.; van Rossum-Fikkert, S. E.; Wyman, C. Conformational flexibility and oligomerization of BRCA2 regions induced by RAD51 interaction. *Nucleic Acids Res* **2020**, 48 (17), 9649-9659. DOI: 10.1093/nar/gkaa648.
- (274) Tan, Z. W.; Tee, W. V.; Samsudin, F.; Guarnera, E.; Bond, P. J.; Berezovsky, I. N. Allosteric perspective on the mutability and druggability of the SARS-CoV-2 Spike protein. *Structure* **2022**, 30 (4), 590-607 e594. DOI: 10.1016/j.str.2021.12.011.
- (275) Tee, W. V.; Guarnera, E.; Berezovsky, I. N. Reversing allosteric communication: From detecting allosteric sites to inducing and tuning targeted allosteric response. *PLoS Comput Biol* **2018**, 14 (6), e1006228. DOI: 10.1371/journal.pcbi.1006228.
- (276) Tresguerres, M.; Clifford, A. M.; Harter, T. S.; Roa, J. N.; Thies, A. B.; Yee, D. P.; Brauner, C. J. Evolutionary links between intra- and extracellular acid-base regulation in fish and other aquatic animals. *J Exp Zool A Ecol Integr Physiol* **2020**, 333 (6), 449-465. DOI: 10.1002/jez.2367.
- (277) Bozovic, O.; Jankovic, B.; Hamm, P. Sensing the allosteric force. *Nat Commun* **2020**, 11 (1), 5841. DOI: 10.1038/s41467-020-19689-7.
- (278) Lockbaum, G. J.; Leidner, F.; Rusere, L. N.; Henes, M.; Kosovrasti, K.; Nachum, G. S.; Nalivaika, E. A.; Ali, A.; Yilmaz, N. K.; Schiffer, C. A. Structural Adaptation of Darunavir Analogues against Primary Mutations in HIV-1 Protease. *ACS Infect Dis* **2019**, 5 (2), 316-325. DOI: 10.1021/acsinfectdis.8b00336.
- (279) Duda, T.; Yadav, P.; Sharma, R. K. Allosteric modification, the primary ATP activation mechanism of atrial natriuretic factor receptor guanylate cyclase. *Biochemistry* **2011**, 50 (7), 1213-1225. DOI: 10.1021/bi1018978.
- (280) Munack, S.; Roderer, K.; Okvist, M.; Kamarauskaite, J.; Sasso, S.; van Eerde, A.; Kast, P.; Kregel, U. Remote Control by Inter-Enzyme Allostery: A Novel Paradigm for Regulation of the Shikimate Pathway. *J Mol Biol* **2016**, 428 (6), 1237-1255. DOI: 10.1016/j.jmb.2016.01.001.
- (281) Walker, T. E.; Shirzadeh, M.; Sun, H. M.; McCabe, J. W.; Roth, A.; Moghadamchargari, Z.; Clemmer, D. E.; Laganowsky, A.; Rye, H.; Russell, D. H. Temperature Regulates Stability, Ligand Binding (Mg(2+) and ATP), and Stoichiometry of GroEL-GroES Complexes. *J Am Chem Soc* **2022**, 144 (6), 2667-2678. DOI: 10.1021/jacs.1c11341.
- (282) Raab, M.; Sanhaji, M.; Pietsch, L.; Bequignon, I.; Herbrand, A. K.; Suss, E.; Gande, S. L.; Caspar, B.; Kudlinzki, D.; Saxena, K.; et al. Modulation of the Allosteric Communication between the Polo-Box Domain and the Catalytic Domain in Plk1 by Small Compounds. *ACS Chem Biol* **2018**, 13 (8), 1921-1931. DOI: 10.1021/acscchembio.7b01078.
- (283) Stamogiannos, A.; Maben, Z.; Papakyriakou, A.; Mpakali, A.; Kokkala, P.; Georgiadis, D.; Stern, L. J.; Stratikos, E. Critical Role of Interdomain Interactions in the Conformational Change and Catalytic Mechanism of Endoplasmic Reticulum Aminopeptidase 1. *Biochemistry* **2017**, 56 (10), 1546-1558. DOI: 10.1021/acs.biochem.6b01170.
- (284) Zaremba, M.; Sasnauskas, G.; Urbanke, C.; Siksnys, V. Allosteric communication network in the tetrameric restriction endonuclease Bse634I. *J Mol Biol* **2006**, 363 (4), 800-812. DOI: 10.1016/j.jmb.2006.08.050.
- (285) Ferrari, D.; Niks, D.; Yang, L. H.; Miles, E. W.; Dunn, M. F. Allosteric communication in the tryptophan synthase bienzyme complex: roles of the beta-subunit aspartate 305-arginine 141 salt bridge. *Biochemistry* **2003**, 42 (25), 7807-7818. DOI: 10.1021/bi034291a.
- (286) Appadurai, R.; Senapati, S. Dynamical Network of HIV-1 Protease Mutants Reveals the Mechanism of Drug Resistance and Unhindered Activity. *Biochemistry* **2016**, 55 (10), 1529-1540. DOI: 10.1021/acs.biochem.5b00946.

- (287) Chen, J.; Peng, C.; Wang, J.; Zhu, W. Exploring molecular mechanism of allosteric inhibitor to relieve drug resistance of multiple mutations in HIV-1 protease by enhanced conformational sampling. *Proteins* **2018**, *86* (12), 1294-1305. DOI: 10.1002/prot.25610.
- (288) Meng, X. M.; Hu, W. J.; Mu, Y. G.; Sheng, X. H. Effect of allosteric molecules on structure and drug affinity of HIV-1 protease by molecular dynamics simulations. *J Mol Graph Model* **2016**, *70*, 153-162. DOI: 10.1016/j.jmglm.2016.09.015.
- (289) Ung, P. M.; Dunbar, J. B., Jr.; Gestwicki, J. E.; Carlson, H. A. An allosteric modulator of HIV-1 protease shows equipotent inhibition of wild-type and drug-resistant proteases. *J Med Chem* **2014**, *57* (15), 6468-6478. DOI: 10.1021/jm5008352.
- (290) Imhof, P. A Networks Approach to Modeling Enzymatic Reactions. *Methods Enzymol* **2016**, *578*, 249-271. DOI: 10.1016/bs.mie.2016.05.025.
- (291) Perez-Hernandez, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noe, F. Identification of slow molecular order parameters for Markov model construction. *J Chem Phys* **2013**, *139* (1), 015102. DOI: 10.1063/1.4811489.
- (292) Ingebrigtsen, T.; Heilmann, O. J.; Toxvaerd, S.; Dyre, J. C. Time reversible molecular dynamics algorithms with holonomic bond constraints in the NPH and NPT ensembles using molecular scaling. *J Chem Phys* **2010**, *132* (15), 154106. DOI: 10.1063/1.3363609.
- (293) *caret: Classification and Regression Training. R package version 6.0-88.* ; 2021. <https://CRAN.R-project.org/package=caret> (accessed).
- (294) *infotheo: Information-Theoretic Measures. R package version 1.2.0.* ; 2014. <https://CRAN.R-project.org/package=infotheo> (accessed).
- (295) Comelli, N. C.; Duchowicz, P. R.; Castro, E. A. QSAR models for thiophene and imidazopyridine derivatives inhibitors of the Polo-Like Kinase 1. *Eur J Pharm Sci* **2014**, *62*, 171-179. DOI: 10.1016/j.ejps.2014.05.029.
- (296) Fatemi, M. H.; Heidari, A.; Gharaghani, S. QSAR prediction of HIV-1 protease inhibitory activities using docking derived molecular descriptors. *J Theor Biol* **2015**, *369*, 13-22. DOI: 10.1016/j.jtbi.2015.01.008.
- (297) Gupta, A. K.; Sabarwal, N.; Agrawal, Y. P.; Prachand, S.; Jain, S. Insights through AM1 calculations into the structural requirement of 3,4,6-substituted-2-quinolone analogs towards FMS kinase inhibitory activity. *Eur J Med Chem* **2010**, *45* (8), 3472-3479. DOI: 10.1016/j.ejmech.2010.05.001.
- (298) Raskevicius, V.; Kairys, V. Predicting Isoform-specific Binding Selectivities of Benzensulfonamides Using QSAR and 3D-QSAR. *Curr Comput Aided Drug Des* **2017**, *13* (1), 75-83. DOI: 10.2174/1573409912666161129153820.
- (299) Turkmenoglu, B.; Guzel, Y. Molecular docking and 4D-QSAR studies of metastatic cancer inhibitor thiazoles. *Comput Biol Chem* **2018**, *76*, 327-337. DOI: 10.1016/j.compbiolchem.2018.07.003.
- (300) Redkar, S.; Mondal, S.; Joseph, A.; Hareesha, K. S. A Machine Learning Approach for Drug-target Interaction Prediction using Wrapper Feature Selection and Class Balancing. *Mol Inform* **2020**, *39* (5), e1900062. DOI: 10.1002/minf.201900062.
- (301) Tianqi Chen, T. H., Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho.; Kailong Chen, R. M., Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin.; Yifeng Geng, Y. L. a. J. Y. xgboost: Extreme Gradient Boosting. R package version 1.6.0.1. . **2022**.
- (302) Bischl B, L. M., Kotthoff L, Schiffner J, Richter J, Studerus E, Casalicchio G, Jones Z. mlr: Machine Learning in R. *_Journal of Machine Learning Research_*, *17*(170), 1-5. . **2016**.