

# Role of local environments in stabilizing protein structures

A thesis submitted in partial fulfillment of the requirement  
of the degree of Doctor of Philosophy

By

**Tejashree Rajaram Kanitkar**

**Registration No. - 20173568**



Department of Biology

Indian Institute of Science Education and Research Pune

India - 411008

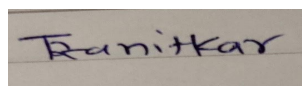
*To*

*Aai, Bhagya, Baba and Sushant*

# DECLARATION

I declare that this written submission represents my idea in my own words and where others' ideas have been included; I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date: 17/07/2023



Tejashree Rajaram Kanitkar  
20173568  
IISER Pune

# CERTIFICATE

Certified that the work incorporated in the thesis titled 'Role of local environments in stabilizing protein structures', submitted by Tejashree Rajaram Kanitkar was carried out by the candidate, under my supervision. The work presented here or any part of it has not been included in any other thesis submitted previously for the award of any degree or diploma from any other university or institution.

Date: 17/07/2023



Dr. M. S. Madhusudhan  
Professor  
IISER Pune

# Acknowledgments

I would like to begin by thanking my mentor and guide Dr. M. S. Madhusudhan for his invaluable guidance, support and encouragement. Whatever little skills and critical thinking I have developed over the years are because of him. I also really want to thank him for the incredible amount of patience he has shown towards me during our discussion sessions.

I would like to thank my lab members. Neelesh was incredibly helpful with his troubleshooting ideas about Packpred. I had interesting and fun discussions (about research and otherwise) with Neeladri, Atreyi and Mukundan that made my time at IISER worthwhile. I am thankful to Atreyi and Neeladri for reviewing my thesis. I want to thank Gulzar for all his help with technical things like setting up systems and programs and also for our informal discussions. Another set of people that I enjoyed hanging out with are Kaustubh, Avadhoot and Kritika. I would also like to thank Parichit, Sanjana, Yogendra, Golding, Ankit, Akash and Swastik for helping me from time to time. I learned a lot from each and every one of you. Thanks for being a part of this journey.

I want to thank Shraddha for being there for me and helping me out in various aspects of work and life as well. She was incredibly supportive and helpful. I am also thankful to my batchmates, especially, Rajeshwari, Shruti, Sandra, and Arjun for being the weird human beings that they are. It was fun spending time with you guys.

I especially want to mention Parvathi for being a really good friend and support system to me. I also want to thank Urmila Kulkarni Kale mam and Asim Auti sir for helping and advising me about my career.

Research in IISER was made faster and easier by the HPCs. I thank Nisha Kaurkure and Neeta deo for their help. I would also like to thank the Bio office and academic office for their prompt responses and help that made our lives easier.

Finally, I want to thank my mother, father and Bhagyashree for being strong pillars of support throughout my life and for being there for me during my good and bad times. Aai, especially, has

been a great driving force and a big source of encouragement in my life. I also want to thank my partner Sushant for being supportive and helping me in various ways. This would not have been possible without you guys.

# Contents

<b>DECLARATION.....</b>	<b>3</b>
<b>CERTIFICATE.....</b>	<b>4</b>
<b>Acknowledgments.....</b>	<b>5</b>
<b>1. Synopsis.....</b>	<b>11</b>
<b>2. Thesis organization.....</b>	<b>14</b>
Chapter 1: Overview.....	14
Chapter 2: Introduction to techniques.....	14
Chapter 3: Molecular mechanism of Class A GPCR activation.....	14
Chapter 4: Packpred: Predicting the functional effect of missense mutations.....	14
Chapter 5: Predicting and Designing therapeutics against the Nipah Virus.....	15
Chapter 6: Conclusion and future prospects.....	15
Chapter 7: Appendix.....	15
<b>Chapter 2.....</b>	<b>16</b>
<b>Introduction to computational techniques.....</b>	<b>16</b>
2.1 Cliques.....	16
2.2 3D least squares fit for geometric comparison of protein structures.....	17
2.3 Classification assessment measures.....	18
2.4 Small molecule binding pocket prediction and molecular docking.....	19
2.5 MD simulations.....	20
2.6 Statistical potentials.....	22
2.7 Sequence alignments and Shannon entropy.....	22
2.8 r - groups.....	23
2.9 Depth and residue depth.....	25
<b>Chapter 3.....</b>	<b>26</b>

<b>Molecular mechanism of Class A GPCR activation.....</b>	<b>26</b>
<b>3.1 Introduction.....</b>	<b>27</b>
<b>3.2 Materials and Methods.....</b>	<b>29</b>
3.2.1 Library of GPCR structures.....	29
3.2.2 Algorithm to identify conserved cliques.....	31
3.2.3 Testing the predictions using molecular dynamics simulations.....	37
<b>3.3 Results.....</b>	<b>38</b>
3.3.1 Designing algorithm for detecting structurally and functionally important local environments.....	38
3.3.2 Inactive state analysis.....	39
3.3.3 Active state analysis.....	46
3.3.4 The r-groups that undergo conformational changes during activation.....	53
3.3.5 Newly formed contacts for stabilizing the active state.....	55
3.3.6 Cliques important for structural stability.....	56
3.3.7 Validation using data from literature.....	57
3.3.8 Testing using Molecular dynamics simulations.....	64
<b>3.4 Discussion.....</b>	<b>67</b>
<b>Chapter 4.....</b>	<b>71</b>
<b>Packpred: Predicting the functional effect of missense mutations.....</b>	<b>71</b>
4.1 Introduction.....	72
<b>4.2. Materials and methods.....</b>	<b>73</b>
4.2.1. Data sets.....	73
4.2.1.1 Statistical potential data set.....	73
4.2.1.2 Saturation mutagenesis data sets.....	74
4.2.1.3 Missense3D data set.....	74
4.2.2 Structural and Sequential features.....	75
4.2.2.1 Residue depth.....	75
4.2.2.2 Cliques of amino acid residues.....	75



4.2.2.3 Statistical potential and residue clique score.....	76
4.2.2.4 Shannon Entropy.....	76
4.2.2.5 FADHM scores.....	77
4.2.2.6 The Packpred score for mutations.....	77
<b>4.3 Results.....</b>	<b>78</b>
4.3.1 Training and testing Packpred score.....	78
4.3.2 Analysis of the predictions on the Missense3D data set.....	80
4.3.3 Meta predictions.....	83
4.3.4 Rank ordering the degree of phenotypic change by mutations.....	85
4.3.5 Assessing robustness of Packpred.....	85
<b>4.4 Discussions.....</b>	<b>86</b>
<b>Chapter 5.....</b>	<b>90</b>
<b>Designing putative inhibitory small molecules against the Nipah virus proteins.....</b>	<b>90</b>
<b>5.1 Introduction.....</b>	<b>91</b>
<b>5.2 Methods.....</b>	<b>92</b>
5.2.1 Prediction of putative small molecules that can bind to NiV proteins:.....	92
5.2.2. Accessing the stability of small molecules against the NiV proteins:.....	93
<b>5.3 Results.....</b>	<b>94</b>
5.3.1 Prediction of putative small molecules that can bind to NiV proteins:.....	94
5.3.2 Computational prediction of the stability of the protein-inhibitor complexes:.....	112
<b>5.4 Discussions.....</b>	<b>119</b>
<b>Chapter 6.....</b>	<b>121</b>
<b>Conclusions and future prospects.....</b>	<b>121</b>
<b>6.1 GPCR.....</b>	<b>121</b>
6.1.1 GPCR summary.....	121
6.1.2 Suggestions and future directions.....	122
<b>6.2 Packpred.....</b>	<b>124</b>
6.2.1 Packpred summary.....	124

6.2.2 Suggestions and future directions.....	125
<b>6.3 Nipah.....</b>	<b>126</b>
6.3.1 Nipah summary.....	126
6.3.2 Suggestions and future directions.....	126
<b>6.4 Concluding remarks.....</b>	<b>128</b>
<b>Chapter 7.....</b>	<b>129</b>
<b>Appendix.....</b>	<b>129</b>
<b>7.1 Progression of algorithm development to detect 3D conserved local environments in GPCRs.</b>	<b>129</b>
7.1.1 Versions of algorithm development and refinement.....	129
<b>7.2 Preliminary analysis that we performed on the data presented chapter 4 Packpred.....</b>	<b>131</b>
<b>Publications.....</b>	<b>158</b>
<b>References.....</b>	<b>159</b>
<b>Copyright forms.....</b>	<b>171</b>

# Chapter 1

## Overview

### 1. Synopsis

Proteins could be represented as local interacting regions also called as local environments. The local environment includes the amino acid and its surrounding residues that interact through hydrogen bonds, pi - pi stacking, cation - pi stacking, Van Der Waals, electrostatics etc. In this thesis, the surrounding residues are identified based on a distance cut-off from the amino acid of interest. The distance cutoff usually ranges from 4Å to 10Å. Some of these local environments can contribute to structural stability of a protein, and/or its functionality. When such local environments are perturbed, it could cause loss of structure and/or function, or it could have no effect on the structure and function of protein. Hence it is necessary to study these local environments and how they affect proteins. The information will assist in the design of proteins with desirable properties. It will also aid in detecting/diagnosing disease conditions, etc.

In this thesis, we studied the local environments using three different examples. In the first one, we used an example of G - protein coupled receptors (GPCRs) to study how perturbation of the local environment caused by ligand binding (agonist) leads to activation of the receptor. GPCRs are signal transmitting molecules that are embedded in the lipid bilayer. Structurally, they consist of seven transmembrane helices. An activating stimuli leads to activation of a GPCR, triggering an intracellular signaling cascade in response to the stimuli. The receptor activation occurs through a series of conformational changes, eventually leading to a movement of the transmembrane helix 6. We wanted to study if the molecular rearrangements that lead to the receptor activation are conserved across different types of GPCRs. If not, in how many distinct ways can they be activated? To address this question, we analyzed 48 Class A GPCR structures in the inactive and 15 structures in the active state separately, to find conserved 3D structural motifs, also called as cliques. Based on the conservation of chemical and geometrical properties of the cliques, we predicted 18 cliques that are important for GPCRs. By comparing the conservations from the inactive state and the active states with each other, we attempted to segregate the

conserved regions that are important for structure and those important for function. We found that in 10 cliques at least 1 and a maximum of 5 r-groups maintain their position during activation, indicating their importance in structural stability. The r-group representation is a novel way of representing similarities in substructures of the amino acids. They are geometric centers of a cluster of covalently connected atoms that form parts of amino acids. 20 amino acids are represented as 16 r-groups. Different types of amino acids can share one r-group. We also found that 15 cliques are either partially or completely disrupted during the activation process, indicating their role in the transition of the receptor from the inactive to the active state. Next, We validated our findings using already reported experimental data. In cases where experimental data were unavailable, we performed molecular dynamics simulations to validate our findings. The results also suggest that GPCRs could be modular in nature. Meaning, that the region responsible for activation can be coupled with any other domain that binds to a ligand of our interest. The modularity could be used as a principle to design novel GPCRs. This information can also be used to design desired mutations while allowing the GPCR to be still functional.

The second example that we studied is where the local environment is perturbed because of the mutations. We tried to predict the effect that a mutation would have on the structure and function of a protein. Towards this, we developed a tool, Packpred, that predicts if a missense mutation would have a deleterious or neutral effect on the structure and function of a protein. To predict the effect, Packpred uses both sequence and structure based features. The sequence based feature is Shannon entropy that quantifies the evolutionarily conserved residues. This allows us to identify structurally and functionally important residues. The structure based features are the packpred statistical potential and FADHM substitution matrices. The statistical potential calculates a log odds ratio of a local environment being observed in the PDB database to that of it occurring merely by chance. It evaluates the likelihood of occurrence of a local environment at a particular depth level. The final feature that Packpred uses is the FADHM substitution matrix. This substitution matrix indicates the probability of substitution of amino acids at different depths. We used a linear combination of the three feature scores by training Packpred on a saturation lysozyme dataset of ~2000 mutations. We tested Packpred on the CcdB saturation mutagenesis dataset containing ~1500 mutations and the Missense3D dataset containing ~4000 mutations. We compared Packpred with 6 other state-of-the-art methods and showed that Packpred outperformed all the other methods in the Missense3D dataset. We also found that although Packpred is good at classifying the effect of the mutation as neutral or deleterious, it is unable to correctly rank order

the mutations based on the severity of the mutation on the phenotype.

In the third example, we tried to predict small molecules that bind to a given local environment (binding pockets). Specifically, we tried to predict the small molecules that can bind to the various binding pockets of 5 proteins (Glycoprotein, Nucleoprotein, Phosphoprotein, Fusion protein, and Matrix protein) from the Nipah virus proteome. We predicted the binding pockets and their druggability using various tools. We then used these predicted binding pockets to dock the small molecule ligands. We selected a 70% non-redundant set of 22,685 clean drug like molecules from the ZINC database. We then performed a virtual screening of these small molecules with the predicted binding pockets of Nipah proteins using docking softwares, Autodock4 and DOCK6.8. We selected 150 best scoring docked complexes for each binding pocket from both Autodock4 and DOCK6.8 and shortlisted the ligands that are common to both runs. Further, to increase the confidence of the predictions, we also calculated the RMSD of the shortlisted small molecule ligands between the autodock and dock complexes. We used the top 5 poses from each run to calculate the RMSD. The final list of selected molecules had a RMSD better than 1.5Å. We then performed MD simulations to assess the stability of the small molecule protein complexes using AMBER99SB-ILDN force field and in some cases CHARMM27 force field. We also calculated the binding energies of the protein-ligand complexes using MM/PBSA software.

To summarize, we studied local environments in proteins and the different effects that a perturbed local environment can have on structure and function of proteins. This study can act as a starting point for other studies that require protein design in different biological contexts.

## 2. Thesis organization

The thesis is organized in six chapters that allows for an easy read as follows.

### Chapter 1: Overview

This chapter explains the need of studying protein structures as local environments. We use various examples like GPCRs that undergo conformational changes to perform its functions. We also use examples of T4 lysozyme and CcdB to study the effect of the perturbed local environment on the structure and function of proteins.

### Chapter 2: Introduction to techniques

This chapter introduces and summarizes various techniques/software that we have used throughout this thesis.

### Chapter 3: Molecular mechanism of Class A GPCR activation

In this chapter we attempted to address if all the Class A GPCRs undergo activation by a universal mechanism, and if not, then in how many ways does it happen? Towards this, we identified local environments that are conserved across different structures, indicating their importance. We further segregated these conserved regions as important either for structure or function or both. We validated this data using either literature or by running Molecular dynamics simulations.

### Chapter 4: Packpred: Predicting the functional effect of missense mutations

In this chapter, we designed a tool, Packpred, that predicts the effect of mutations on the structure and function of a protein. We trained Packpred on ~2000 mutations of T4 lysozyme saturation mutagenesis dataset and tested on ~6000 mutations belonging to the CcdB saturation mutagenesis dataset and the Missense3D dataset. We compared our performance with 6 other state-of-the-art methods and showed that Packpred outperforms others in the Missense3D dataset.

## Chapter 5: Predicting and Designing therapeutics against the Nipah Virus

In this chapter, we predicted and designed putative small molecules and peptides that would bind to and inhibit different Nipah proteins. Particularly, for predicting small molecules, we predicted binding pockets using the DEPTH web server and performed docking using Autodock4 and DOCK6.8. We then shortlist ligands by having a consensus of top 50 best scoring ligand-protein complexes from the 2 docking software that also have similarities in the poses of the docked ligand. We also performed MD simulations to assess the stability of the ligand-protein complexes.

## Chapter 6: Conclusion and future prospects

This chapter summarizes the findings of this thesis and also discusses the new ideas that could be implemented for refining the existing algorithms.

## Chapter 7: Appendix

This chapter summarizes the progression of algorithms described in chapter 3 and 4.

## Chapter 2

# Introduction to computational techniques

## 2.1 Cliques

Proteins are made of amino acids that are connected by peptide bonds. The sequence of amino acids that forms a protein is called its primary structure. Hydrogen bonds between the backbone of the amino-acids often lead to formation of secondary structure. The protein eventually folds into a three-dimensional(3D) structure called its tertiary structure. The primary structure determines the tertiary structure of a protein[1]. Proteins that have similar sequences of amino acids fold into similar tertiary structures[2]. The 3D structures of proteins can be represented as clusters of amino acids that interact with each other, known as clique. Cliques are local regions of proteins that can be used to study and characterize the local interactions. Thus, a protein contains many such cliques. Characterization of cliques can assist in the design of novel proteins, thermostable proteins, proteins that have desired characteristics, etc. Previously, cliques have been used to perform structure based alignments[3, 4], derive statistical potentials for evaluating models of proteins[5], etc. Although the definition of clique might vary depending on the necessity of the study, the basic idea of studying the local environment remains constant. Some of the ways a clique can be defined are:

- 1.1. All the residues that lie within a distance cutoff of a central residue (Figure 1A)
- 1.2. All the residues whose pairwise distances are less than a distance cutoff (Figure 1B)
- 1.3. Nearest N residues from the central one - where N is a pre decided constant (Figure 1C)

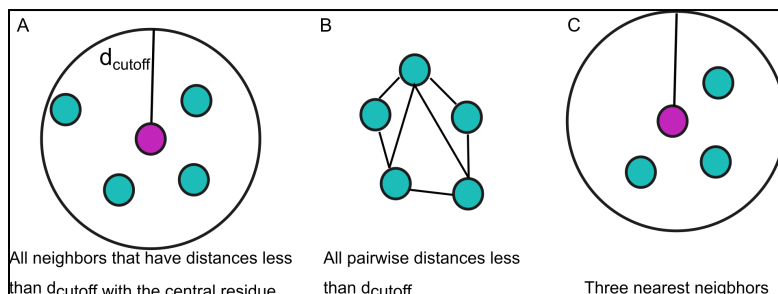
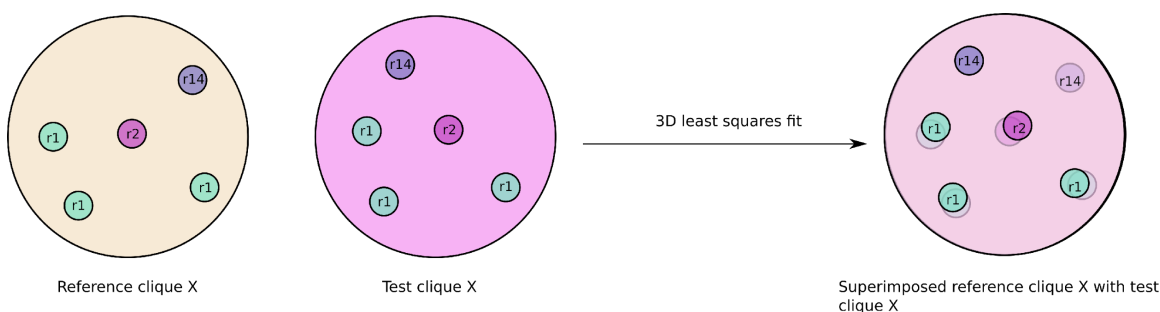


Figure 1: Different ways of defining cliques. The central residue/atom is represented as a pink circle. All neighbors are represented as teal coloured circles.



## 2.2 3D least squares fit for geometric comparison of protein structures

The 3D least squares fit is used to compare the geometry of structures/sets of 3D points. It has been used in various software that perform structure based alignments of biomolecules. Specifically, it helps evaluate if the points in both the sets are arranged in a similar fashion. It does so by minimizing the distances between two sets of points. 3D least squares fit needs equivalences between the points as input. Based on the equivalences, it transforms the coordinates to get the least root mean square deviation (RMSD)[6]. The transformed coordinates provide the superimposition of the set of points/structures (Figure 2). The similarity between the two sets of points can be quantified and evaluated using the RMSD of superimposition and/or the number of superimposed points (that are close to each other as identified by a predetermined distance cut-off). A lower RMSD and a higher number of superimposed points indicate similarity between the sets. Superimposition of identical sets will have an RMSD of 0.00 with all the points matched. Algorithms by Diamond[7], Kearsley[8], Kabasch[9] and many others can be used to perform the structural superimposition. In this study, we have used a 3D least squares fit algorithm by Kearsleys to find geometric similarities between two sets of points.



*Figure 2: Reference clique X and test clique X are two sets of points in 3D space. Each of them consists of 5 points indicated by smaller circles. The colors of smaller circles indicate their equivalences. For instance, the purple circle (r2) in reference clique X is equivalent to the purple circle (r2) in test clique X and so on. Based on these equivalences, 3D least squares fit performs superimposition by minimizing the distances between the equivalences giving the least root mean square deviation.*

## 2.3 Classification assessment measures

A classification scheme classifies data points into different categories. A binary classifier

classifies data points into two categories. The performance of the binary classifier can be assessed using different metrics. Following are some of the widely used assessment metrics.

1.4. Confusion matrix: Confusion matrix categorizes the binary classification predictions into four categories:

1.4.1. True positive (TP): The data points that have a positive label and are predicted as positive

1.4.2. True negative(TN): The data points that have a negative label and are predicted as negative

1.4.3. False positive(FP): The data points that have a negative label but are predicted as positive

1.4.4. False negative(FN): The data points that have a positive label but are predicted as negative

1.5. Matthews correlation coefficient (MCC): MCC takes into account TP, TN, FP and FN to evaluate the performance of the classifier. It is a balanced measure that is unaffected by different proportions of positively and negatively labeled data points. It is given by,

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

1.6. Sensitivity (Recall): It is a measure of how many positively labeled data points were correctly predicted by the classifier. It is given by,

$$Sensitivity = \frac{TP}{TP + FN}$$

1.7. Specificity: It is a measure of how many negatively labeled data points were correctly predicted by the classifier. It is given by,

$$Specificity = \frac{TN}{TN + FP}$$

1.8. Precision: Precision indicates the number of correctly predicted positives out of all the positively predicted data points. It is given by,

$$Precision = \frac{TP}{TP + FP}$$

1.9. F1 score: F1 is a harmonic mean of precision and recall. It is given by,

$$f1 = 2 * \left( \frac{Precision * Recall}{Precision + Recall} \right)$$

- 1.10. Accuracy: Accuracy indicates the total number of correctly predicted data points by a classifier. It is given by,

$$Accuracy = \frac{TP + TN}{TN + TP + FP + FN}$$

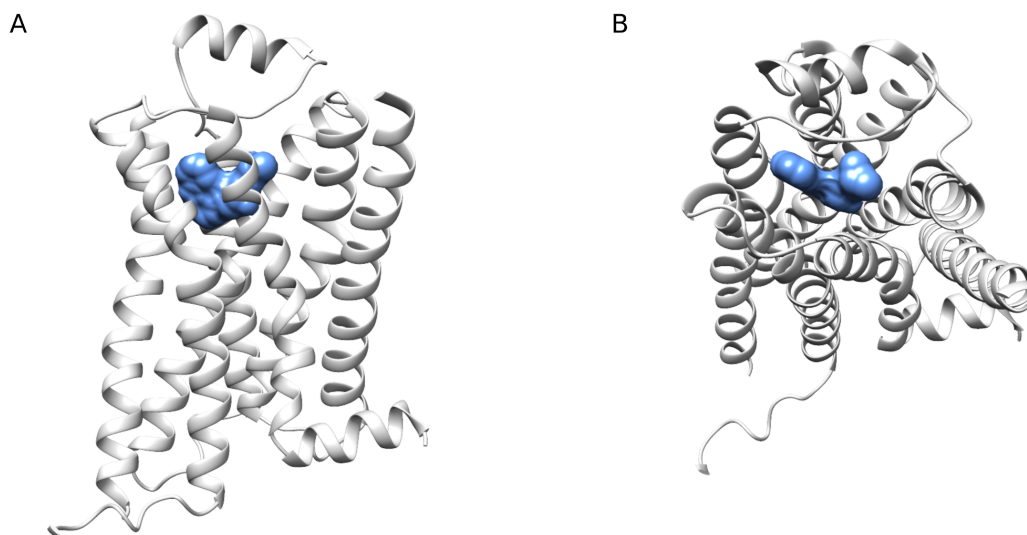
## 2.4 Small molecule binding pocket prediction and molecular docking

The region on the protein where a ligand binds is called a ligand binding pocket (Figure 3). Predicting binding pockets on the surface of a protein is the first step in designing a small molecule ligand that binds to it. Predicting binding pockets is complicated because of lack of knowledge of its partner ligand, induced fit changes that occur upon binding, etc. Hence, various binding pocket prediction tools exploit different sequence and structure based properties of the proteins. For example, methods like LIGSITE[10], SURFNET[11] make use of geometry of the protein while Consurf[12], FINDSITE[13], 3DLigandSite[14] make use of evolutionary information and templates available in the PDB. Other methods make use of accessible solvent area as well. Once the binding pockets are predicted, we can model its interaction with a ligand through a method called molecular docking. All the docking methods usually consist of two broad steps. The first step samples various poses of the ligand. The sampling is followed by a second step of scoring each pose based on the contacts it makes with the protein[15]. Various docking softwares sample the ligand poses differently. Some of the algorithms used are geometry based, fragment based and stochastic searches. The scoring schemes are used to identify the correct binding pose by ranking it better than the incorrect poses. They make use of physics based methods, empirical methods or knowledge based methods to score the complexes[15].

The predicted binding pockets can also be used to computationally screen a library of small molecule ligands to predict those that can potentially bind to the pocket[16, 17]. This process is called virtual screening. Thousands of small molecule ligands can be computationally screened against a binding pocket. These processes are fast and computationally inexpensive. The screening can be performed using molecular docking software like AutoDock[18], DOCK[19], SwissDock[20], and many others[21].

In some cases, the ligand that binds to a protein of interest is known but its binding pose is

unknown. Docking can be used to predict the correct binding pose of the ligand in the binding pocket. In the absence of known binding pockets, docking could also be performed on the entire surface of the protein. This is known as blind docking. Instead of limiting the sampling search to a binding pocket, blind docking uses the entire surface of the protein for sampling, making it slower than local docking (where a binding pocket is specified)[22, 23]. Another feature that the docking softwares offer is the use of rigid or flexible docking. In rigid docking, the entire protein is considered as a non-flexible entity that does not undergo any movement. However, in reality, proteins undergo events like induced fit upon binding to its partner. In such cases, flexible docking can be performed, where movement of a set of user specified residues is allowed and accounted for during the docking exercise[22, 23].



*Figure 3: A) Ligand represented in blue surface bound to a GPCR represented as light gray ribbon, indicating ligand binding pocket (PDB: 2RH1). B) Top view of the ligand bound region*

## 2.5 MD simulations

Proteins are flexible biomolecules that undergo molecular motions in solvent. These movements also assist in forming complexes with other molecules through the induced fit mechanism. These movements in proteins can be studied via computer simulations using Molecular dynamics (MD) simulations[24]. Given a 3D structure of a molecule, the simulations calculate the subsequent movements of the atoms as a function of time using Newton's laws of motion. The position and velocity of each atom in the molecule is updated, giving a trajectory of movement of atoms over

time. In nature, forces like electrostatics, Van Der Waals etc. govern the movements of the atoms. In the MD simulations, these forces/interactions are accounted for in the form of a force field. A force field is an approximation that quantifies all such interactions using data from various experiments/data sources. Some examples of force fields are CHARMM, AMBER, and the OPLS. A generalized formula for force field is given by,

$$\begin{aligned}
 U(r) = & \sum_{bonds} k_b (b - b_0)^2 + \sum_{angles} k_\theta (\theta - \theta_0)^2 + \\
 & \sum_{dihedrals} k_\chi (1 + \cos(n\chi - \delta)) + \\
 & \sum_{vdW, i \neq j} \epsilon_{ij} \left[ \left( \frac{R_{min,ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{R_{min,ij}}{r_{ij}} \right)^6 \right] + \sum_{elec, i \neq j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}
 \end{aligned}$$

Where, U is interatomic potential energy that is represented as a forcefield,

$k_b$ ,  $k_\theta$  and  $k_\chi$  are force constants for bonds lengths, bond angles and dihedral angles,

$b_0$  and  $\theta_0$  are equilibrium values for the bond length and valence angle between atoms,

$n$  is the dihedral multiplicity,

$\delta$  is the dihedral angle phase,

vdW are the van der Waals forces,

elec are the electrostatic interactions [25]

MD simulations have been used to study conformational changes that occur in proteins, refine 3D structure models of biomolecules, study interactions between biomolecules etc. Simulations of different types of biological systems such as a single protein in an aqueous solvent, protein embedded in the membrane, small molecule-protein complexes, protein-protein complexes, protein-DNA complexes can be simulated using software like GROMACS[26], NAMD[27], AMBER[28] and many others. Additionally, the simulations can also be performed at varied levels of resolution like all atoms, or clubbing several atoms together by coarse graining, or a hybrid of molecular and quantum mechanics. The MD simulations require high computational power[29]. With the advances in computational hardware like GPUs and supercomputers, performing simulations has become fast and less expensive.

## 2.6 Statistical potentials

Protein structure prediction methods have immensely aided in getting insights into the structure of a protein and also its putative function. These methods can be broadly classified into 3 classes, namely, homology based methods, threading and de novo predictions[30]. The quality of the models predicted by these methods depends on the data used to generate them. It is hence essential to quantify their correctness. Statistical potentials, also known as knowledge based potentials, are a category of scoring functions that can be used to gauge the correctness of the models[29]. The potentials are derived from a database, such as Protein Data Bank (PDB), hence called knowledge based potentials. They often capture some features of proteins based on the chemistry and physics of the interactions from the database[31]. The features are quantified as ratio of observed by expected frequencies/probabilities as formulated by Sippl[32]. The observed data is extracted from known protein structures and the expected value is formulated such that it indicates the occurrence of a feature by chance. Some of the widely used statistical potentials include DOPE[5], GOAP[33], ROTAS[34], and ProSA[32]. These statistical potentials differ from one another in the terms of protein representation and the spatial feature that is used. Protein can be represented in various ways, some of them are as an all-atom system or by its C<sup>α</sup> atom or by side chain centroid. Some spatial features that these potentials capture are distances, torsion angles or angles etc. The protein representation and the spatial feature in combination with the reference state affect the performance/accuracy of the statistical potentials.

## 2.7 Sequence alignments and Shannon entropy

Pairwise sequence alignments attempt to identify the similarities, and conserved regions between two sequences. The alignments could be performed on protein, DNA and RNA sequences. Alignments can be of two types, the global alignment, where the sequences are compared to each other end-to-end[35]. In local alignment, local regions of similarities in the provided sequences are identified. Local alignments are typically performed using tools such as BLAST[36], PSI-BLAST[37], a slightly different version of BLAST, allows searching of the local regions of similarities with an entire database, allowing identification of distant homologs. The alignment of multiple sequences is called multiple sequence alignment. These alignments are an invaluable tool to identify evolutionary relationships. The evolutionarily conserved positions in the sequence show higher conservation and less variation in the multiple sequence alignment. The variation can be quantified by Shannon entropy[38], given by,

$$H(X) = \sum_{i=1}^n P(X_i) \log_2 P(X_i)$$

Where,  $H(X)$  is Shannon entropy at position  $X$ .  $P(X_i)$  is the probability of occurrence of a particular character ( $i$ ) at position  $X$ ,  $n$  is the number of unique characters that can occur.

In the case of amino acid sequences,  $n = 20$  representing 20 amino acids. The Shannon entropy for amino acids is represented as,

$$H(X) = \sum_{i=1}^{20} P(X_i) \log_2 P(X_i)$$

The Shannon entropy ranges from 0 to 4.32 for the 20 standard amino acids, with 0 being the most conserved position and 4.32 being the least conserved/ highly variable position[39].

## 2.8 r - groups

Typically, proteins are represented as a sequence of amino acids. Depending on the necessity of the study, the amino acids are represented in different ways such as heavy atoms,  $C^\alpha$  atoms, etc. In this study, we have represented the 20 amino acids as r-groups (Figure 3). The r-groups are created by dividing the amino acids into smaller sections that have distinct properties. We used r-groups as the protein family of our interest, GPCRs, are known to have as low as 35% sequence identity amongst themselves. Thus, we wanted to check if only smaller regions of amino acids are participating and responsible for the structure and function of GPCRs. The idea for this representation is that in some cases only subregions of the amino acids may be necessary for the interaction/stability of the clique. We defined 16 such r-groups by clubbing various covalently linked heavy atoms. The r-groups are represented as a centroid of the 3D coordinates of its constituent atoms. This allowed us to represent one r-group as one point in 3D space, and are named as r1, r2, ..., r16 in this study (Figure 3).

The backbone of all amino acids, except proline, is represented as r1. Since proline is the only amino acid whose side chain forms a closed ring with the backbone atoms, it is categorized as the r11 group. Glycine does not have a heavy atom in its sidechain and is represented as r1. Similarly, we defined three r-groups, r14, r15 and r16 for the side chains of the aromatic residues, PHE, TRP and TYR respectively. While r14 is aromatic, the OH of the r16 group gives it a polar nature and hence is categorized as a different group. The r15 group is the indole ring of TRP where the lone pair of N also participates in the aromatic ring, giving it different properties than the PHE

and TYR. We have defined all the 16 r-groups in a similar manner, considering the properties of the clubbed atoms. One thing to note is that these groups are created heuristically and could be defined in various ways.

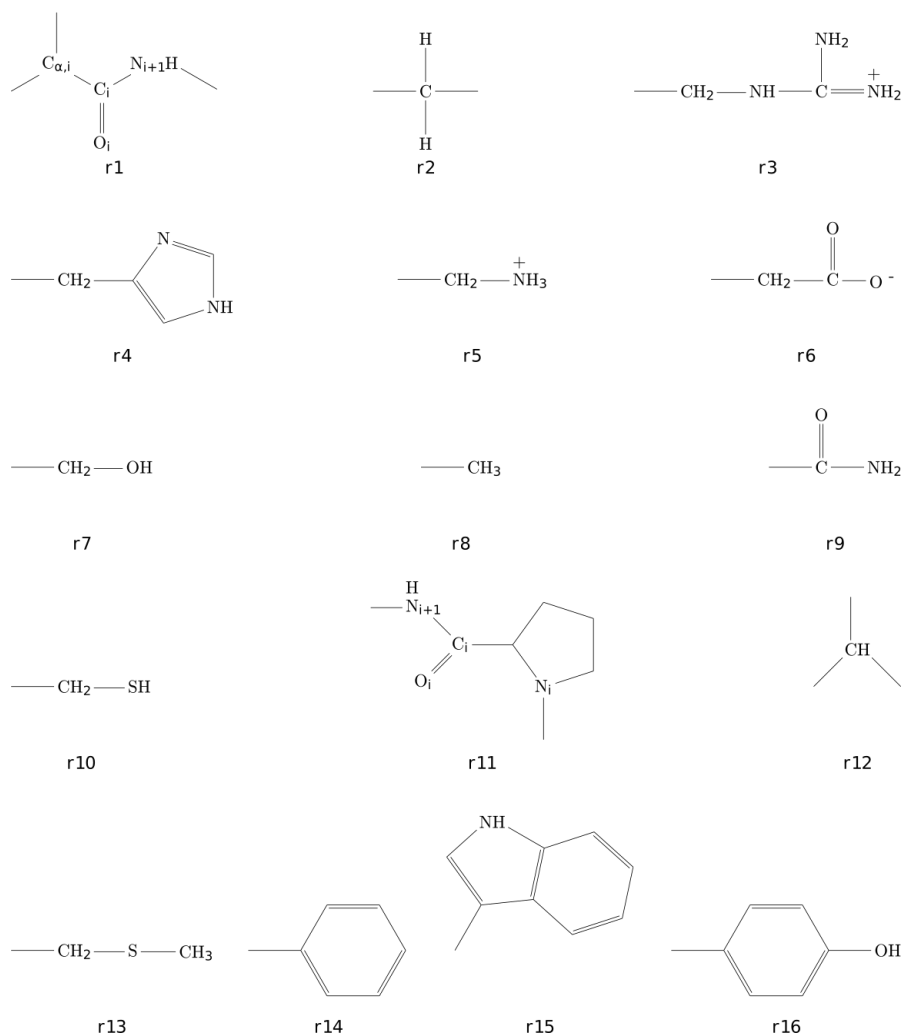


Figure 3: The definition of 16 r-groups used to represent 20 amino acids (Image adapted from Master's thesis of Akash Bahai <http://dr.iiserpune.ac.in:8080/xmlui/handle/123456789/570> )

## 2.9 Depth and residue depth

In an aqueous solution, a protein folds into its tertiary structure such that the hydrophilic amino



acids are exposed to the solvent while the hydrophobic amino acids are buried inside the core of the protein. The burial of the hydrophobic residues occludes their unfavorable interactions with the aqueous solvent. Thus, amino acids within a protein are placed at different distances from the solvent.

Traditionally, the extent to which the protein is buried/accessible to solvent is quantified using a metric called as accessible surface area[40]. A more stratified quantification of the burial of the residues is given by the definition of residue depth[41, 42]. For aqueous proteins, depth is defined as the distance between the protein atom and its nearest water molecule in the bulk solvent. The bulk solvent excludes water molecules that are trapped inside the protein cavities. A bulk solvent has at least 4 (or user specified value of water molecules) other neighboring water molecules within its hydration sphere of 1.5 layers. Non-bulk solvent molecules have less than 4 neighboring waters in its hydration shell. Residue depth is calculated as an average of depths of all the atoms of amino acids.

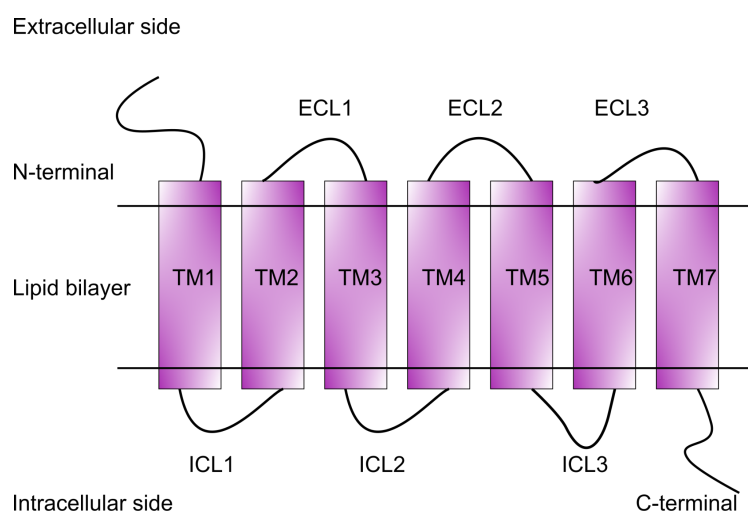
## Chapter 3

### Molecular mechanism of Class A GPCR activation

- Creation of a dataset of Class A GPCR structures
- Representation of the structures as r-groups
- Identification of structurally conserved 3D motifs
- Prediction of functionally important structural 3D motifs
- Validation
  - Data from the literature
  - MD simulations

### 3.1 Introduction

G protein-coupled receptors (GPCRs) are a class of signal transducing proteins found in the membranes of cells[43]. These receptors are activated by a variety of stimuli, including binding of various extracellular ligands, such as hormones, neurotransmitters, sensory stimuli, and many others. The activation triggers and initiates intracellular signaling pathways that result in the production of various second messenger molecules, such as cyclic adenosine monophosphate (cAMP)[44], which can then transmit the signal further into the cell. Because of this function, GPCRs are involved in a wide range of physiological processes, including the immune response[45, 46], sensory perception[47, 48], and regulation of various metabolic pathways[49]. This also makes them attractive targets for the development of drugs to treat a wide range of diseases, including cardiovascular diseases, neurological disorders, and metabolic disorders. 103 from a total of 403 GPCRs are targeted by ~40% marketed drugs, leaving a large number of them yet to be targeted[50].



*Figure 1: A schematic of GPCR architecture consisting of seven transmembrane helices (TM1 - 7) that are connected on the extracellular side by 3 extracellular loops (ECL1 - 3) and on the intracellular side by 3 intracellular loops (ICL1 - 3). The N-terminal of the receptor is exposed outside the cell while the C-terminal lies inside the cell.*

Structurally, GPCRs are made of seven transmembrane(TM) helices that traverse through the lipid bilayer (Figure 1). These helices are connected by three loops on the intracellular and

extracellular sides. The N-terminal lies outside the cell while C-terminal lies inside the cell. The activating ligand usually binds on the extracellular side of the GPCR, at a site known as the orthogonal ligand binding site. The binding stabilizes the receptor in a state that can interact with other molecules like the G-proteins or the arrestins on the intracellular side, known as the active state. In addition to this conserved architecture, GPCRs also have conserved motifs like NPxxY, DRY, PIF, CWxP, and Na<sup>+</sup> pocket[51–53]. All these motifs are functionally important. The NPxxY and DRY motifs are located on the TM 7 and TM3 respectively, near the cytoplasmic side and are known to be important for activation of receptors like rhodopsin, B2AR, oxytocin and A2AR[54–57]. The PIF is a structural motif formed by three hydrophobic residues contributed by TM5, TM5 and TM6. This triad of residues is observed near the bottom of the ligand binding pocket. Repacking of these triad residues is observed in many active state GPCRs. The ‘W’ of the CWxP motif on TM6 is particularly important for activation of GPCRs. It is called a toggle residue as it changes its rotameric state upon receptor activation[58]. Na<sup>+</sup> ion in the Na<sup>+</sup> pocket is coordinated by TM2, TM3, and TM7 in inactive conformations of most GPCRs. The ion is displaced from its pocket and is not seen in the active state conformations[52].

Based on the sequence identity of the transmembrane region[59], GPCRs are classified into 6 classes, A to F[60]. Class A, also called the Rhodopsin-like family, is the largest family of GPCRs. The family includes receptors that are activated by small molecules such as hormones, neurotransmitters, and sensory stimuli. Class B GPCRs, also known as secretin and adhesion family, include receptors that are activated by peptides. Examples of Class B receptors include the corticotropin-releasing hormone receptor and the glucagon receptors[61]. Class C GPCRs are activated by the neurotransmitter glutamate. Some examples of Class C receptors are the  $\gamma$ -aminobutyric acid<sub>B</sub> receptors (GABA<sub>B</sub> receptors) and the Ca<sup>2+</sup>-sensing receptors[62]. Class D GPCRs are pheromone receptors that are exclusively found in fungi[63]. Class E GPCRs include receptors that are activated by cAMP. Class F GPCRs include the frizzled receptors. Of the six classes, classes A, B, C and F occur in vertebrates.

There are two main conformations that GPCRs can adopt: the inactive state and the active state[43, 64]. In the inactive state, the receptor does not transmit any signals and does not alter the activity of any signaling pathways. GPCRs are stabilized in another conformation, the active state, by an activating stimuli or an agonist. The activation causes TM6 of GPCRs (Class A) to move outward by  $\sim 14\text{\AA}$ , creating a pocket for interacting with various intracellular proteins,

initiating intracellular signaling pathways. There are 6 Class A GPCRs that have X-ray structures in the inactive as well as the active state, enabling the detailed study of their activation. These 6 receptors are bovine rhodopsin (bRho)[65, 66],  $\beta_2$ -adrenergic receptor ( $\beta_2$ AR)[67, 68], M2 muscarinic receptor (M2R)[69, 70],  $\mu$ -opioid receptor ( $\mu$ OR)[71, 72], adenosine  $A_{2A}$  receptor ( $A_{2A}$ R)[73, 74] and  $\kappa$ -opioid receptor ( $\kappa$ -OR)[75]. However the information is limited as the activation process of a vast majority of GPCRs is still unknown. In this study, we investigated if the GPCR activation process including the conformational changes responsible for transitioning the receptor from inactive to active state are similar across different GPCRs, and if not, how many different ways GPCRs are activated. To address this question, we computationally analyzed Class A GPCR structures to find conserved 3D local regions (3D motifs) of geometric and chemical similarities. The conserved 3D motifs, when put together, form a continuous pathway running from the bottom of the ligand binding pocket to the intracellular/cytoplasmic region of the receptor. We validated our findings using data from literature and by performing molecular dynamics simulations.

## 3.2 Materials and Methods

### 3.2.1 Library of GPCR structures

We retrieved 789 experimentally resolved GPCR structures from the GPCRdb[76] on 22/08/2022 (Table 1). From this dataset, we selected one X-ray structure with the best resolution per gene of Class A GPCR (as per Uniprot) to remove redundancy, leading to 50 structures of the inactive state and 15 structures of the active state (Table 1). 2 of the 50 inactive state structures (PDBID: 7B6W and 6YVR) were left out of the analysis as they were not available in the PDB file format. The final library consisted of 48 inactive and 15 active state structures. All the selected structures had a resolution better than 3Å.

Table 1: (1) List of all inactive state Class A GPCR structures used in this study. (2) List of all active state Class A GPCR structures used in this study. (3) List of all GPCR PDB structures retrieved from the GPCRdb.

Sr no	Dataset	PDB ID
1	Inactive state structures (Class A proteins only)	4IAR, 7WC9, 6BQH, 5NM4, 6ZFF, 5ZKC, 4U15, 5DSG, 6OL9, 7B6W, 6KUX, 6KUW, 4BVN, 6PS2, 4ZUD, 5VBL, 6I9K, 6C1R, 7F8Y, 6GPX, 5UIW, 6QZH, 5LWE, 5U09, 5ZTY, 3ODU, 6CM4, 3PBL, 5WIU, 6IGK, 7F83, 7BR3, 6LI0, 4Z36, 7K15, 6ME2, 6ME6, 6HLP, 5ZBQ, 7DDZ, 6YVR, 4N6H, 4DJH, 4DKL, 5DHH, 1U19, 6TOS, 5WQC, 7M8W
2	Active state structures (Class A proteins only)	5TUD, 6BQG, 5WF5, 6H7N, 4LDE, 6OS2, 5UNF, 5XRA, 6LW5, 4XES, 6PT2, 5C1M, 5DYS, 6M9T, 4XT1
3	789 structures retrieved from GPCRdb (across all classes of GPCRs)	7U2L, 7WU2, 7WU4, 7WU5, 7WU3, 7SF8, 7WUJ, 7SF7, 7WQ3, 7WQ4, 7SBF, 7SCG, 7EZC, 7EJA, 7EJ0, 7EJK, 7EJ8, 7WVU, 7WVW, 7WVX, 7WVY, 7WVW, 7RBT, 7RA3, 7RGP, 7RG9, 7TUZ, 7TUY, 7EJX, 7VBH, 7T6B, 7TYI, 7TYN, 7TYX, 7T6T, 7T6S, 7T6V, 7T6U, 7TYW, 7TYH, 7TZF, 7TYL, 7TYF, 7TYO, 7TYY, 7VL8, 7VL9, 7VLA, 7WI8, 7WIH, 7WI6, 7VKT, 7RYC, 7T10, 7T11, 7PYR, 7PX4, 7FIY, 7VAB, 7FIM, 7VBI, 7V35, 7VGZ, 7VGY, 7VH0, 7FIN, 7VGX, 7JNI, 7TD0, 7TD2, 7TD1, 7TD3, 7TD4, 7WC7, 7WC6, 7WC5, 7WC9, 7WC4, 7WC8, 7RKM, 7RKF, 7RKN, 7SIN, 7SIL, 7SIM, 7W2Z, 7F83, 7B6W, 7LLY, 7LLL, 7S3I, 7S1M, 7WF7, 7EO2, 7EO4, 7F8X, 7VUH, 7VUI, 7VUJ, 7VUG, 7VOD, 7VOE, 7NA8, 7NA7, 7P00, 7P02, 7EWL, 7SHE, 7EWP, 7EWR, 7SHE, 7VDM, 7VV3, 7VDL, 7VUZ, 7VUY, 7VV4, 7VV6, 7VDH, 7VV0, 7VV5, 7V3Z, 7PIV, 7PIU, 7S8M, 7S8O, 7S8L, 7S8N, 7S8P, 7F55, 7F53, 7F54, 7F58, 7RMH, 7RMG, 7RMI, 7V9M, 7EIB, 7F2O, 7F8U, 7F8Y, 7F8V, 7F8W, 6ZFF, 6ZG9, 6ZG4, 7RTB, 7FIG, 7FIH, 7FII, 7FIJ, 7EW7, 7EVZ, 7EVY, 7EW0, 7EW2, 7EW3, 7EW4, 7EW1, 7E6T, 7E6U, 7DGD, 7DGE, 7LD4, 7LD3, 7RM5, 7FD9, 7FD8, 7P2L, 7F4I, 7F4F, 7F4D, 7F4H, 7EZK, 7EZH, 7EZM, 7M8W, 7F9Z, 7F9Y, 7DB6, 7F16, 7DUR, 7EVM, 7EXD, 7DH5, 7EVW, 7DTY, 7KI0, 7KI1, 7DW9, 7F1T, 7F1S, 7F1R, 7F1Q, 7DUQ, 7E14, 7MTS, 7MTQ, 7MTR, 7MTA, 7MT8, 7MTB, 7MT9, 7JHJ, 7M3F, 7M3E, 7M3G, 7M3J, 7O7F, 7EPA, 7E9G, 7EPE, 7EPB, 7EPF, 7EPD, 7E9H, 7EPC, 7DD6, 7DD7, 7DD5, 7C4S, 7BB6, 7BB7, 7MBX, 7MBY, 7KH0, 7EB2, 7CX3, 7CX2, 7CX4, 7AUE, 7E32, 7E2Y, 7E2X, 7E2Z, 7E33, 7JOZ, 7ARO, 7DFL, 7DTT, 7DTU, 7DTV, 7DTW, 7CMU, 7CMV, 7CKX, 7LJC, 7CKY, 7LJD, 7CKZ, 7CKW, 7CRH, 7KNT, 7KNU, 7JV5, 7JVP, 7JVQ, 7JVR, 7K15, 6YVR, 6Z4V, 6ZIN, 6Z4S, 6ZA8, 6Z66, 6Z4Q, 6Z8N, 7L1U, 7L1V, 7D76, 7D77, 7DDZ, 7LCK, 7LCI, 7LCJ, 7LOR, 7L0P, 7L0Q, 7L0S, 7DFP, 7DHR, 7DHI, 7D68, 7AD3, 7BVQ, 7BTS, 7BU6, 7BU7, 6LPK, 6LPL, 6LPJ, 6WQA, 7CZ5, 6XOX, 7D7M, 7CA5, 7CA3, 7CUM, 7D3S, 7BR3, 6XBM, 6XBJ, 6XBK, 6XBL, 6WH4, 6WGT, 6WHA, 6ZDV, 6ZDR, 6Z10, 6X18, 6X1A, 6X19, 7CFN, 7CFM, 7JJO, 6LFL, 6LFM, 6LFO, 7BW0, 6VN7, 7C2E, 6WW2, 6KO5, 6WPW, 6WZG, 6WI9, 7BZ2, 6TPK, 7C61, 7C6A, 6VCB, 6S0Q, 6S0L, 6V9S, 6WJC, 6W2Y, 6W2X, 7C7Q, 6WIV, 7C7S, 6PGS, 6PH7, 6WWZ, 6PEL, 6TKO, 6VMS, 6UO9, 6UO8, 6VJM, 6UOA, 6WHC, 6W25, 6K42, 6K41, 6UUS, 6UVA, 6LMK, 6LML, 6OBA, 6UUN, 6LW5, 6LN2, 6VI4, 6KP6, 6M1H, 6M1I, 6LPB, 6LUQ, 6UIN, 6OMM, 6LI0, 6LI1, 6LI3, 6LI2, 6UP7, 6OS2, 6OS0, 6OS1, 6KPG, 6KPC,

	6KPF, 6PT0, 6PB0, 6PB1, 6LRY, 6NWE, 6P9X, 6P9Y, 6KNM, 6JOD, 6TOT, 6TOS, 6ORV, 6TP3, 6TQ9, 6TP4, 6TQ7, 6TOD, 6TQ4, 6TO7, 6TP6, 6TQ6, 6TPJ, 6TPN, 6TPG, 6OL9, 6RZ8, 6RZ9, 6RZ6, 6RZ7, 6PT2, 6PT3, 6KUY, 6KUX, 6KUW, 6IQL, 6PWC, 6NI3, 6PS7, 6PS0, 6PS4, 6PS1, 6PS5, 6PRZ, 6PS3, 6PS2, 6PS6, 6KJV, 6KK1, 6KK7, 6PS8, 6JZH, 6RZ5, 6RZ4, 6KQI, 6QZH, 6OFJ, 6IBB, 6RNK, 6OYA, 6OY9, 6K1Q, 6OSA, 6OS9, 6QNO, 6I9K, 6O3C, 6GT3, 6N48, 6OT0, 6E67, 6OIJ, 6OIK, 6MH8, 6ME2, 6ME3, 6ME5, 6ME4, 6ME7, 6ME9, 6ME6, 6ME8, 6NBH, 6NBF, 6NBI, 6J21, 6J20, 6A94, 6A93, 6DO1, 6N4B, 5ZTY, 6NIY, 6N51, 6N52, 6HLO, 6HLL, 6HLP, 6IBL, 6GPS, 6GPX, 6IIU, 6IIV, 6MEO, 6MET, 6E59, 6M9T, 6AK3, 5YHL, 5YWY, 5ZHP, 5ZK3, 5ZKC, 5ZK8, 5YC8, 5ZKB, 6IGK, 6IGL, 6FJ3, 6MXT, 6AKX, 6AKY, 6H7L, 6H7J, 6H7O, 6H7M, 6H7N, 6FUF, 6D27, 6D26, 6E3Y, 6DRZ, 6DRY, 6DS0, 6DRX, 6BD4, 5XJM, 6G79, 6D9H, 6CMO, 5ZKQ, 5ZKP, 6DDE, 6DDF, 5WB1, 5WB2, 6C1R, 6C1Q, 6D35, 6D32, 6GDG, 5KW2, 5ZBH, 5ZBQ, 6FK8, 6FKC, 6FK9, 6FKD, 6FK6, 6FKA, 6FK7, 6FKB, 6CM4, 6FFH, 6FFI, 5WF6, 5WF5, 6B3J, 6BQG, 6BQH, 5V54, 5OLV, 5OLO, 5OM4, 5OLG, 5OLZ, 5OLH, 5OM1, 5YQZ, 6B73, 6AQF, 5O9H, 5X33, 5VRA, 5WK7, 5WS3, 5WQC, 5WIV, 5WIU, 5NM2, 5NM4, 5NLX, 5X7D, 5XPR, 5X93, 5XSZ, 5W0P, 5TUD, 5N2S, 5MZJ, 5MZP, 5N2R, 5XR8, 5XRA, 5UIW, 5NX2, 5TZR, 5TZY, 5JTB, 5VBL, 5UVI, 5VAI, 5VEW, 5XF1, 5XEZ, 5V56, 5V57, 5VEX, 5UZ7, 5NDZ, 5NJ6, 5NDD, 5UNF, 5UNG, 5UNH, 5TE3, 5TE5, 5UEN, 5UIG, 5TVN, 5T04, 5T1A, 5LWE, 5U09, 5TGZ, 5K2C, 5K2A, 5K2B, 5K2D, 5GL1, 5GLH, 5D6L, 5DYS, 5EN0, 5G53, 5L7I, 5L7D, 5JQH, 5IUA, 5IU7, 5IU8, 5IU4, 5IUB, 4Z9G, 5EE7, 5DGY, 5DSG, 5CXV, 4ZJ8, 4ZJC, 5D5B, 5D5A, 5F8U, 4X1H, 5DHH, 5DHG, 4ZUD, 5A8E, 5CGD, 5CGC, 5C1M, 4XES, 4XEE, 4ZWJ, 4WW3, 4Z36, 4Z35, 4Z34, 4YAY, 4UG2, 4UHR, 4XNV, 4XNW, 4XT1, 4XT3, 4RWS, 4RWA, 4RWD, 4S0V, 4U16, 4U15, 4U14, 4PXF, 4QKX, 4QIN, 4QIM, 4PHU, 4O09, 4PY0, 4PXZ, 4BVN, 4NTJ, 4OR2, 4O9R, 4BWB, 4BV0, 4BUO, 3ZEV, 4N4W, 4N6H, 4NC3, 4MQS, 4MQT, 4J4Q, 4LDE, 4LDO, 4LDL, 4MBS, 4L6R, 4K5Y, 4BEY, 4BEZ, 4JKV, 3ZPQ, 3ZPR, 4IAR, 4IAQ, 4IB4, 4GPO, 3VW7, 4GBR, 4GRV, 4EYI, 4AMI, 4AMJ, 4EJ4, 4EA3, 3UZC, 3UZA, 4DJH, 4DKL, 4DAJ, 3V2Y, 3V2W, 3VG9, 3VGA, 3UON, 4A4M, 3PWH, 3REY, 3RFM, 3AYM, 3AYN, 3SN6, 3RZE, 2YCY, 2YCX, 2Y CZ, 2YCW, 2YDO, 2YDV, 2Y01, 2X72, 3QAK, 3PXO, 3PQR, 3P0G, 3OAX, 2Y00, 2Y02, 2Y03, 2Y04, 3PDS, 3PBL, 3ODU, 3OE0, 3OE6, 3OE8, 3OE9, 3NY8, 3NY9, 3NYA, 3KJ6, 3EML, 3DQB, 3C9L, 3C9M, 2VT4, 3CAP, 3D4S, 2Z73, 2Z1Y, 2R4R, 2R4S, 2RH1, 2PED, 2J4Y, 2I35, 2I36, 2I37, 2G87, 2HPY, 1U19, 1GZM, 1L9H, 1HZX, 1F88
--	---

### 3.2.2 Algorithm to identify conserved cliques

We attempted to identify the 3D motifs/cliques that are essential for the structure and function of Class A GPCRs. Towards this, we designed an algorithm that identifies the chemical and geometrical similarities between various Class A receptors. Our method then identifies the cliques that are chemically and geometrically conserved in different receptors by taking a consensus. To

identify the consensus of the conserved cliques, we designed an algorithm consisting of 9 steps. These steps are described below.

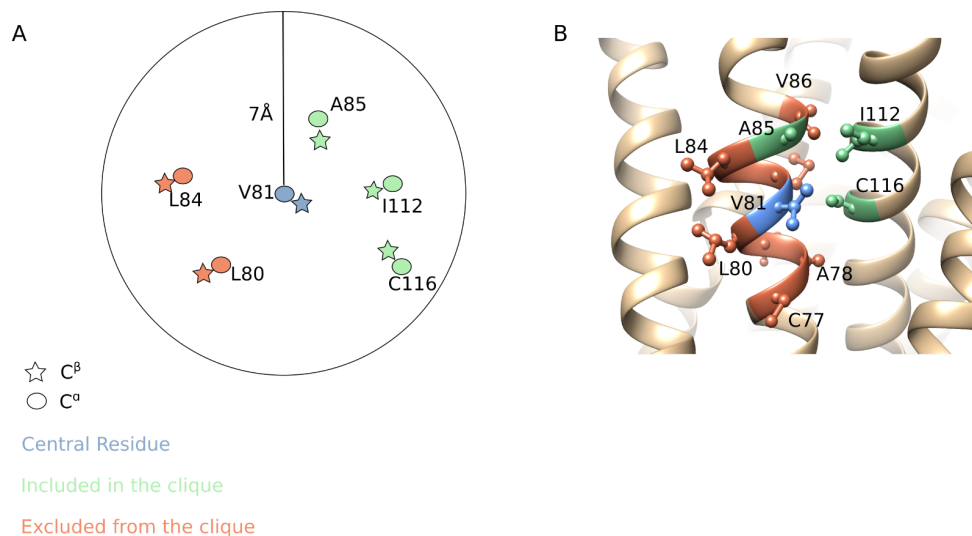
- a. Completion of the structure by filling in missing atoms: For each PDB structure, we selected and completed the receptor chain for any missing atoms using the `complete_pdb` module of `modeller v9.25`[77, 78].
- b. Global superimposition and residue equivalences: We obtained the residue equivalences by superimposing all the 48 inactive structures from our library with an inactive state reference structure using `TMalign`[79]. The reference is a randomly selected Class A GPCR structure from the GPCRdb dataset (leaving out the structures from our library), with which the library structures are superimposed. Similarly, we obtained residue equivalences for active state by superimposing active structures with an active state reference.
- c. Defining local environments as cliques: The residue for which we are constructing a clique is called a central residue. All the residues that are within a distance of 7Å from the C<sup>α</sup> atom of the central residue are called its neighbors. We further apply the following criteria to filter in only those residues whose side chains are pointing towards the central residue, indicating that they are interacting with each other.

$$d(C_{cen}^{\alpha}, C_{neb}^{\alpha}) > d(C_{cen}^{\beta}, C_{neb}^{\beta}) \quad (1)$$

Where, C<sub>cen</sub><sup>α</sup> is the C<sup>α</sup> atom of the central residue, C<sub>neb</sub><sup>α</sup> is the C<sup>α</sup> atom of the neighboring residue, C<sub>cen</sub><sup>β</sup> is the C<sup>β</sup> atom of the central residue, C<sub>neb</sub><sup>β</sup> is the C<sup>β</sup> atom of the neighboring residue and d is the distance between the specified atoms.

A clique consists of at least 3 such residues that satisfy equation 1 (Figure 2). This criterion allowed us to include only the residues that are interacting with each other. We constructed such cliques for all the residues of the reference structure.





*Figure 2: (A) Schematic of clique definition: Residues 85, 112 and 116 form a clique of central residue 81. Residue number 80 and 84 are eliminated from the clique as its C<sup>β</sup> distance with the central residue is greater than their C<sup>α</sup> distances (B) PDB 2RH1 rendered as beige ribbon. Residues 85, 112 and 116 form a clique of central residue 81. Residue number 77, 78, 80, 84 and 86 are eliminated from the clique as its C<sup>β</sup> distance with the central residue is greater than their C<sup>α</sup> distances*

- d. Constructing library cliques: We constructed the cliques for the structures from the library using cliques from step c and residue equivalences from step b.
- e. Representing residues as r-groups: We represented the 20 amino-acids as 16 r-groups (Figure 3). Please refer to section 8 of Chapter 2 for details of the r-groups.

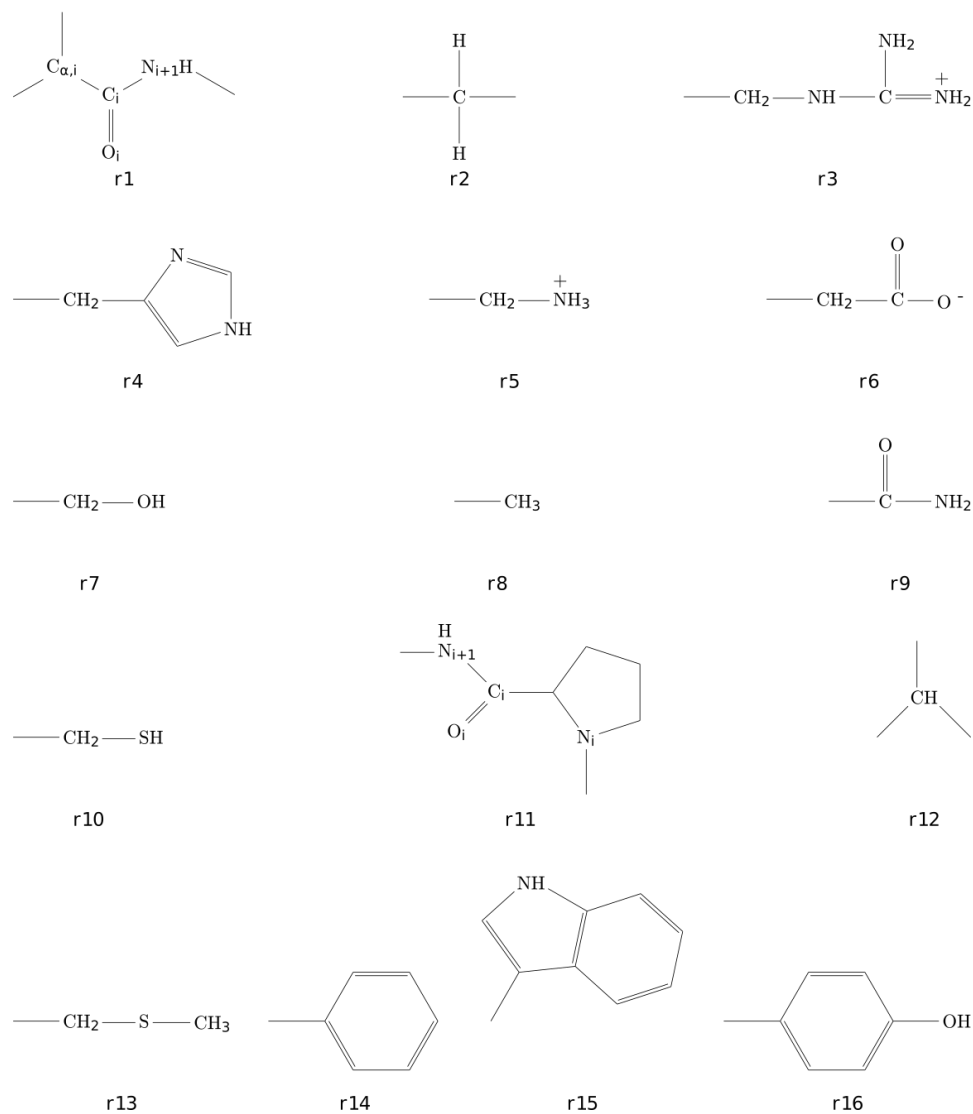
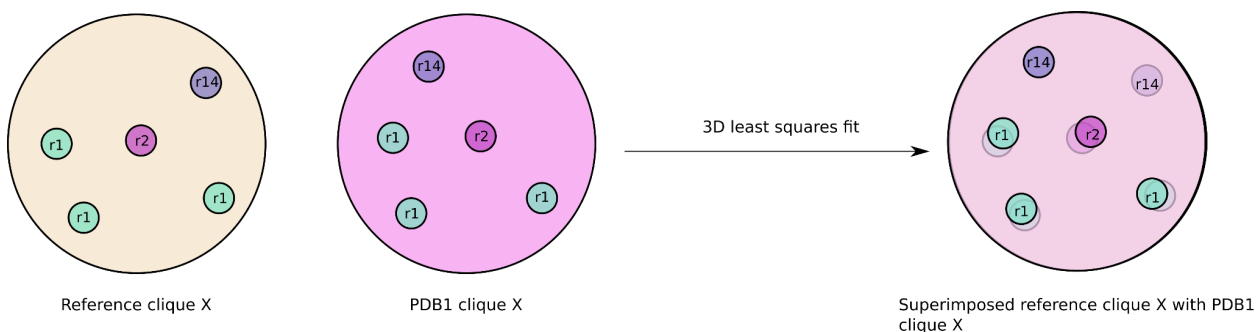


Figure 3: The definition of 16 r-groups used to represent 20 amino acids (Image adapted from Master's thesis of Akash Bahai <http://dr.iiserpune.ac.in:8080/xmlui/handle/123456789/570> )

- f. Local superimposition: Local superimpositions allow us to identify geometrical similarities of the r-group arrangements between two cliques. In this study, we wanted to identify similarities between the reference clique and the library cliques. Hence, we performed 2 rounds of local superimpositions for each reference clique with the library cliques. The first round of the superimposition was performed using the r1 and r11 (backbone groups). This was followed by another round, round 2 of local

superimpositions which aimed at refining or fine tuning the superimpositions from round1. The details of the two rounds of superimpositions are as follows.

- i. Round 1 of local superimposition and local equivalences: For each residue clique in the reference structure, we superimposed its corresponding library clique using 3D least squares fit. Equivalences from step b (backbone r1 and proline r11) were used as input for the 3D least squares fit. Using this superimposition, we found the closest r-group of the library clique to each r-group from the reference clique, giving us local equivalences of round 1. We calculated equivalences for all r-groups except r1 and r11 (the r1 and r11 equivalences established in step b are retained). The closest r-group cannot be farther away than 1.5 Å (Figure 4)



*Figure 4: An example of local superimposition using 3D least squares fit. The reference clique has 3 r1 groups (cyan circles), one r2 group (dark pink circle) and one r14 group (purple circle). PDB1 clique too has three r1, one r2 and one r14. These two cliques are superimposed using 3D least squares fit. The superimposed image shows that three r1 groups and one r2 group found a match in PDB1 with three r1 and one r2 group. The r14 group from the reference does not match with r14 from PDB1 as the distance between them after superimposition is greater than 1.5 Å.*

- ii. Round 2 of local superimposition and local equivalences: Using the equivalences from step b and round 1, we performed another round of local superimpositions. Similar to round 1, we found the closest r-group pairs from the reference clique and the library clique (round 2 equivalences), indicating geometrical similarities. We performed all the further analysis using the equivalences from round 2.
- g. Finding conserved r-groups and conserved cliques: Conserved r-groups are those that have less variation of r-groups at the same geometric location (as identified by round 2 local superimpositions) in various GPCR structures. We imposed 2 criteria to find such

r-groups (Figure 5)

- i. The r-group found an equivalent match in at least 60% of the structures from the library
- ii. The Shannon entropy[38] ( $H(X)$ , Equation 1) of the superimposed r-groups was less than 1.00:

For instance, the last row in Figure 5 indicates the number of matched PDBs for reference r-group r15 contributed by a TRP of residue number 286. It has a match frequency of 41, which is the sum of the row. Its Shannon Entropy is 0.8. Since both, match frequency and Shannon Entropy fulfill the set criteria, the r15 from TRP286 is considered as a conserved r-group.

$$H(X) = \sum_{i=1}^{16} P(X_i) \log_2 P(X_i) \quad \text{Eq (1)}$$

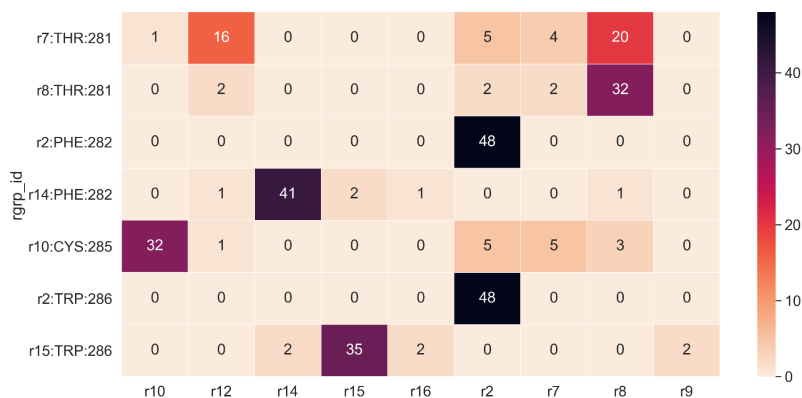


Figure 5: An example of match frequency of all r-groups of clique of residue number 318. This clique has 7 r-groups represented on the Y-axis. All the r-groups that are observed in the library cliques are indicated on the X-axis. The numbers of heatmap indicate the frequency of a r-group on Y-axis matching with an r-group from the library cliques.

A clique that has 3 such conserved r-groups contributed by at least 2 amino acids are called conserved cliques.

- h. Since we were comparing all the structures to one reference structure, the conservation results may change when a different reference structure is used. We repeated steps b-g with 2 additional reference structures selected randomly from the GPCRdb to eliminate

the dependency on one reference structure. We then took a consensus of the conserved r-groups from 3 reference structures. The r-groups that are conserved in the analysis of at least 2 reference structures are called consensus conserved r-groups and the cliques that are formed from these r-groups are called consensus conserved cliques. The consensus conserved r-groups and consensus conserved cliques are also referred to as conservations in this study.

- i. We repeated steps b-h for inactive state structures and active state structures separately. This gave us the consensus r-groups that are important for inactive state and active state. In addition, we also looked at the consensus r-groups that are conserved in inactive state but not in active state indicating that they moved during the process of activation and are hence essential for receptor activation. Further, we also found the consensus r-groups that are conserved in active state but not in inactive state indicating their importance in achieving/maintaining the active state conformation.

### 3.2.3 Testing the predictions using molecular dynamics simulations

We performed Molecular dynamics simulations (MDs) using version 2021.3 of the GROMACS[80][25] package to test the residues that did not have any mutation data in the literature. We performed triplicates of 200ns CHARMM36[81] all atom simulations of  $\beta$ 2AR receptor along with its ligand (PDBID: 2RH1). We inserted  $\beta$ 2AR in a POPC bilayer having 76 molecules in the upper leaf and 75 molecules in the lower leaf, to maintain the XY dimension ratio as 1 (to have identical system size along the X and Y). The orientation of  $\beta$ 2AR in the bilayer was determined using the PPM2.0 server. The POPC bilayer embedded receptor was solvated using a TIP3P rectangular water box of thickness 22.5cm. We neutralized the system by adding 0.15mM KCL. The approximate system size was 80x80x128, containing ~760,00 atoms. We saved energy after every 2ps. We performed a neighbor search using the Verlet cutoff scheme, where the short range Van Der Waals cutoff was set to 1.2nm. We treated electrostatics using the particle-mesh Ewald method[82] and constrained the hydrogen bond lengths using the LINCS[78] method. We then minimized the system for 5000 steps or till the maximum force was less than 1000 kJ/mol/nm. This was followed by heating of the system to 303.15K in an NVT ensemble for 250 ps using a Berendsen thermostat[83]. We stabilized the pressure in an NPT ensemble simulation for 500 ps using a Berendsen barostat. During the equilibration, positional and

dihedral restraint potentials were applied, and their force constants were gradually reduced. We then simulated the systems (NPT) for a maximum of 200 ns where pressure was regulated using the semi-isotropic Parrinello-Rahman barostat[84]. We stored the structures after every 5ns and monitored the temperature, potential energy and kinetic energy during the simulation to check for anomalies.

## 3.3 Results

### 3.3.1 Designing algorithm for detecting structurally and functionally important local environments

We designed an algorithm that identifies 3D structural motifs that are essential for structure and/or function of a protein based on conserved geometrical and chemical properties. To do this, our algorithm represents a PDB structure as r-groups, which are then represented as cliques. We compare these cliques to a reference structure to find geometric and chemical similarities. We perform 3D least squares fit to find geometric similarity between the cliques. Based on the geometric match, we consider each r-group in the reference clique and its matched partner from the 3D least squares fit as a position of a multiple sequence alignment. We then calculate Shannon's Entropy to quantify variation of r-groups at a particular geometric location, enabling us to find chemical similarities. Shannon entropy allows us to capture variations, including events like swapping of r-groups within a clique. We further apply various heuristic cutoffs such as Shannon entropy (has to be less than 1.00) and the match frequency (greater than 60%) to identify the conserved cliques/r-groups. We repeat this process using 3 distinct reference structures (Table 2) to get higher confidence. We selected these reference structures such that they have different resolution and respond to different types of ligand to eliminate bias created because of it. The cliques and their constituent r-groups that are conserved in at least 2 reference structures analyses are considered to be conserved.

Table 2: Details of reference structures used for the active and the inactive state analysis

Reference PDB ID	Resolution(Å)	Type	State
4X1H	2.3	Rhodopsin	Active
7BU6	2.7	Adrenoceptor	Active
4XEE	2.9	Neurotensin	Active
2RH1	2.4	Adrenoceptor	Inactive
4Z36	2.9	Lysophospholipid	Inactive
5ZBH	3	NeuropeptideY	Inactive

We applied this algorithm to both inactive and active state structures of GPCRs. Conservation in each of the states indicates the functional or structural importance of these cliques to a particular state. The transition from the inactive state to the active state is associated with loss of some contacts and formation of some new contacts. To identify the contacts that are lost during activation, we compared the conservation from the inactive state to that of the active state. The r-groups that are conserved in the inactive state but not in the active state are the ones that are lost during activation. These contacts may be crucial for receptor activation. We then compared active state conservation with the inactive state to identify new contacts that allow the receptor to transition to the active state. These contacts could have a role in stabilizing the active state. Conservations that are common to both active and inactive states can mean that they are necessary for structural stability.

The conserved 3D motifs from our analysis span over the membrane embedded region of GPCR and some intracellular region. The conservations connect the well known motifs such as NPxxY, DRY, CWxP and the Na<sup>+</sup> pocket, creating an activation pathway. We do not get any conservation in the extracellular domain that is involved in the binding of the ligand. Our activation pathway starts a little lower to the ligand binding domain and spans through the membrane embedded region to the intracellular region. We believe that the trigger to the activation is different for different GPCRs, but eventually the trigger associated conformational changes converge into a common pathway that activates GPCRs.

### 3.3.2 Inactive state analysis

We analyzed 48 inactive state structures (Table 1) using three inactive state reference structures (Table 2) to find consensus conserved cliques. The reference structures were selected such that they had different resolutions and responded to a different type of ligand. We found 23, 22 and 16 conserved cliques for 3 reference structure analysis respectively (Table 3). 18 cliques

are common to at least 2 reference structures and 10 cliques are common to all 3 reference structures (Table 3). Further, we looked at the r-group conservation and found that 61 are conserved in at least 2 reference structure analysis (consensus conserved r-groups) and 30 are conserved in all the 3 reference structure analysis (Figure 6A). The consensus conserved r-groups are contributed by 27 amino-acid residues. The Ballesteros-Weinstein (BW) numbering[85] of the 27 amino-acids that contribute to the consensus conserved r-groups is '1x50', '1x52', '1x53', '2x42', '2x45', '2x46', '2x47', '2x49', '2x50', '2x51', '3x43', '3x46', '3x50', '4x49', '4x50', '4x52', '4x53', '5x60', '5x61', '6x39', '6x40', '6x44', '6x48', '7x45', '7x49', '7x52', and '7x53'. Of these, the most conserved residues from helix 1, 2, 3 and 4 (as represented by Helix number x50) are also conserved according to our analysis. The conserved r-group of all 4 helices are contributed by a single type of amino-acid, except for the one from helix number 2, that is contributed by 3 distinct types of amino-acids. The most conserved residue of helix 5, 6 and 7 is a proline, which our analysis treats as an equivalent to a backbone group and hence, is not found to be conserved. 18 of these cliques contain r-groups other than r2, r8 and r12.



Table 3: Consensus cliques conserved from the inactive state in reference structure1, reference structure2, reference structure 3 along with the frequency of consensus conserved cliques from the inactive state in the 3 reference structures. The star\_no column indicates the central residue of the clique. ‘Conserved\_in\_inactive\_reference1’ column indicates conserved residue number and the conserved r-group separated by ‘\_’. All the residue numbers follow the numbering of PDB 2RH1.

star_no	conserved_in_inactive_reference1	conserved_in_inactive_reference2_equivalent_to_reference1	conserved_in_inactive_reference3_equivalent_to_reference1	How_many_in_active_references_it_is_conserved
51	['76_r8', '79_r6', '80_r8']	['76_r8', '79_r6', '80_r8']	['76_r8', '79_r6', '322_r2']	3
54	['72_r8', '76_r8', '326_r2', '326_r16']	['76_r8', '326_r2', '326_r16']	['76_r8', '322_r2', '326_r2']	3
64	['58_r8', '61_r2', '69_r2', '69_r9']	NA	NA	1
71	['127_r12', '127_r8', '154_r8']	['127_r12', '127_r2', '130_r6']	NA	2
76	['51_r2', '51_r9', '54_r12', '54_r8']	['51_r2', '51_r9', '54_r8']	['51_r2', '51_r9', '54_r12', '54_r8']	3
79	['51_r2', '51_r9', '322_r2', '322_r9']	['51_r2', '51_r9', '322_r2', '322_r9']	['51_r2', '51_r9', '120_r7', '322_r2']	3
119	['78_r8', '157_r8', '158_r2', '158_r15', '161_r7']	['78_r8', '158_r2', '158_r15', '161_r7']	['74_r7', '78_r8', '157_r8', '158_r2', '158_r15', '161_r7']	3
121	['208_r2', '208_r14', '282_r2', '282_r14']	NA	NA	1
122	['157_r8', '160_r2', '161_r7']	NA	['157_r8', '160_r2', '161_r7']	2

123	['71_r2', '75_r2', '75_r12', '75_r8']	['71_r2', '74_r7', '75_r2', '75_r12', '75_r8', '78_r8']	['71_r2', '74_r7', '75_r2', '75_r12', '157_r8']	3
124	['75_r8', '278_r12', '278_r8', '282_r2', '282_r14']	['75_r2', '75_r8', '278_r12', '278_r8', '282_r2']	['279_r8', '282_r2', '282_r14']	3
132	['221_r2', '222_r12', '222_r8']	NA	['221_r2', '222_r12', '222_r8']	2
208	['121_r8', '286_r2', '286_r15']	NA	NA	1
211	['121_r12', '121_r8', '122_r2']	NA	NA	1
218	['129_r8', '132_r2', '132_r16', '214_r8']	NA	NA	1
222	['131_r2', '132_r2', '132_r16', '135_r8']	NA	['131_r2', '135_r12']	2
278	['124_r2', '124_r12', '124_r8', '322_r2', '322_r9', '325_r8']	['124_r8', '322_r2', '322_r9', '325_r8']	NA	2
282	['120_r7', '124_r2', '124_r12', '124_r8', '318_r2']	['124_r2', '124_r8', '286_r2', '286_r15', '318_r9']	NA	2
286	['208_r2', '318_r2', '318_r9']	['282_r2', '282_r14', '290_r2', '318_r9']	NA	2
318	['281_r8', '282_r2', '282_r14', '286_r2', '286_r15']	['282_r2', '282_r14', '286_r2', '286_r15']	['79_r6', '282_r2', '282_r14', '286_r2', '286_r15']	3
322	['75_r2', '75_r12', '75_r8', '79_r6', '278_r12', '278_r8']	['75_r2', '75_r8', '79_r6', '278_r12', '278_r8']	['51_r2', '51_r9', '54_r8', '75_r2', '76_r8', '79_r6', '278_r8']	3
323	['51_r2', '51_r9', '53_r2', '53_r8', '54_r12', '54_r8']	['51_r2', '51_r9', '53_r2', '54_r12', '54_r8']	['51_r2', '51_r9', '53_r2', '54_r12', '54_r8', '76_r8']	3
325	['277_r8', '278_r12', '278_r8']	['277_r8', '278_r12']	NA	2

	'281_r8']	'278_r8']		
75	NA	['124_r2', '124_r8', '322_r2', '322_r9']	NA	1
127	NA	['71_r2', '278_r12', '278_r8']	NA	1
223	NA	['131_r2', '132_r2', '132_r16', '135_r8']	NA	1
272	NA	['131_r2', '135_r8', '223_r12', '223_r8']	NA	1
275	NA	['127_r12', '127_r8', '131_r2']	NA	1
285	NA	['314_r2', '314_r8', '318_r9', '321_r8']	NA	1
328	NA	['273_r2', '277_r8', '325_r8']	NA	1
68	NA	NA	['127_r12', '127_r8', '130_r6', '131_r2']	1
215	NA	NA	['124_r2', '124_r8', '279_r8']	1
319	NA	NA	['51_r2', '51_r9', '79_r6']	1

We also found that all the consensus conserved r-groups occur in at least 70% of our dataset(Figure 7A).

A majority (~90%) of the consensus conserved r-groups lie within the membrane embedded region of GPCRs. We also get some conservation in the intracellular region. We do not get any conservation in the extracellular region (Figure 8). We also found that the conserved r-groups are contributed by a maximum of 10 and a minimum of 1 amino-acid, highlighting that the r-groups are not just contributed by a highly conserved amino-acid across GPCRs, but are from different amino-acids.

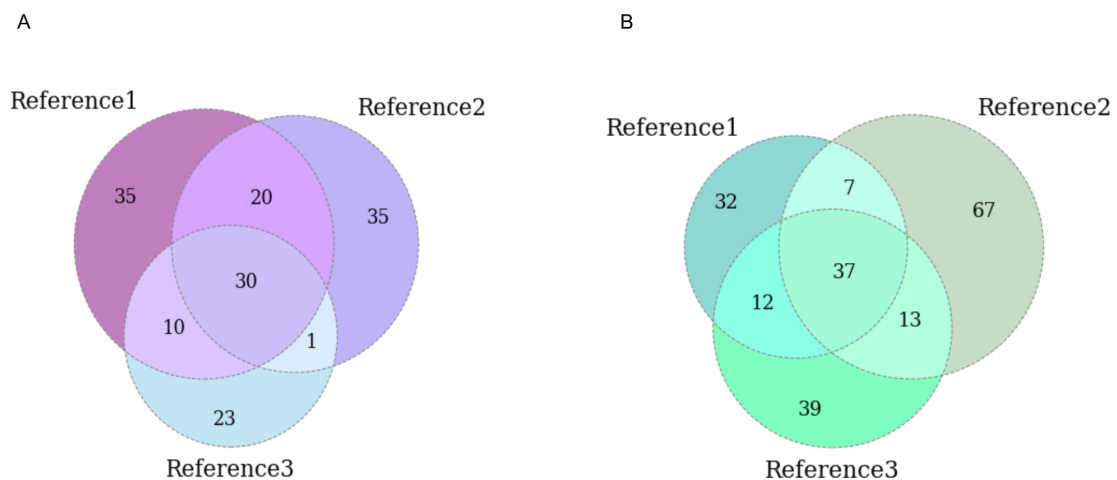


Figure 6: Overlap between conserved r-groups from three reference structures for the inactive state(A) and the active state(B)

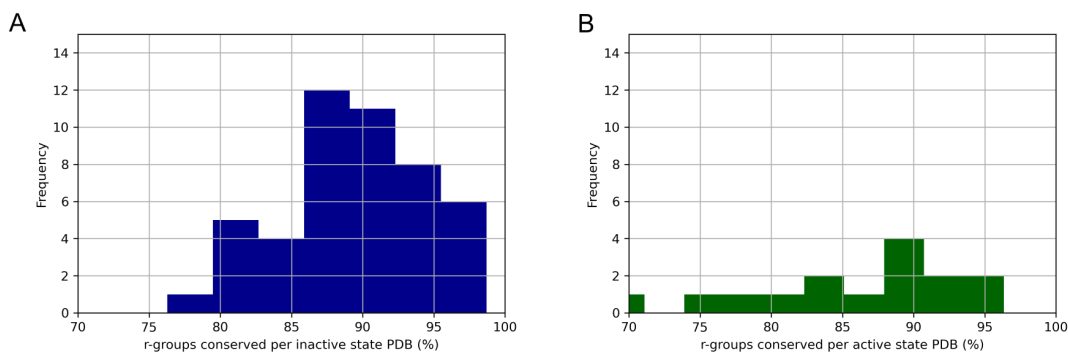


Figure 7: Percentage of PDBs containing Consensus conserved r-groups in the inactive state(A) and the active state(B)

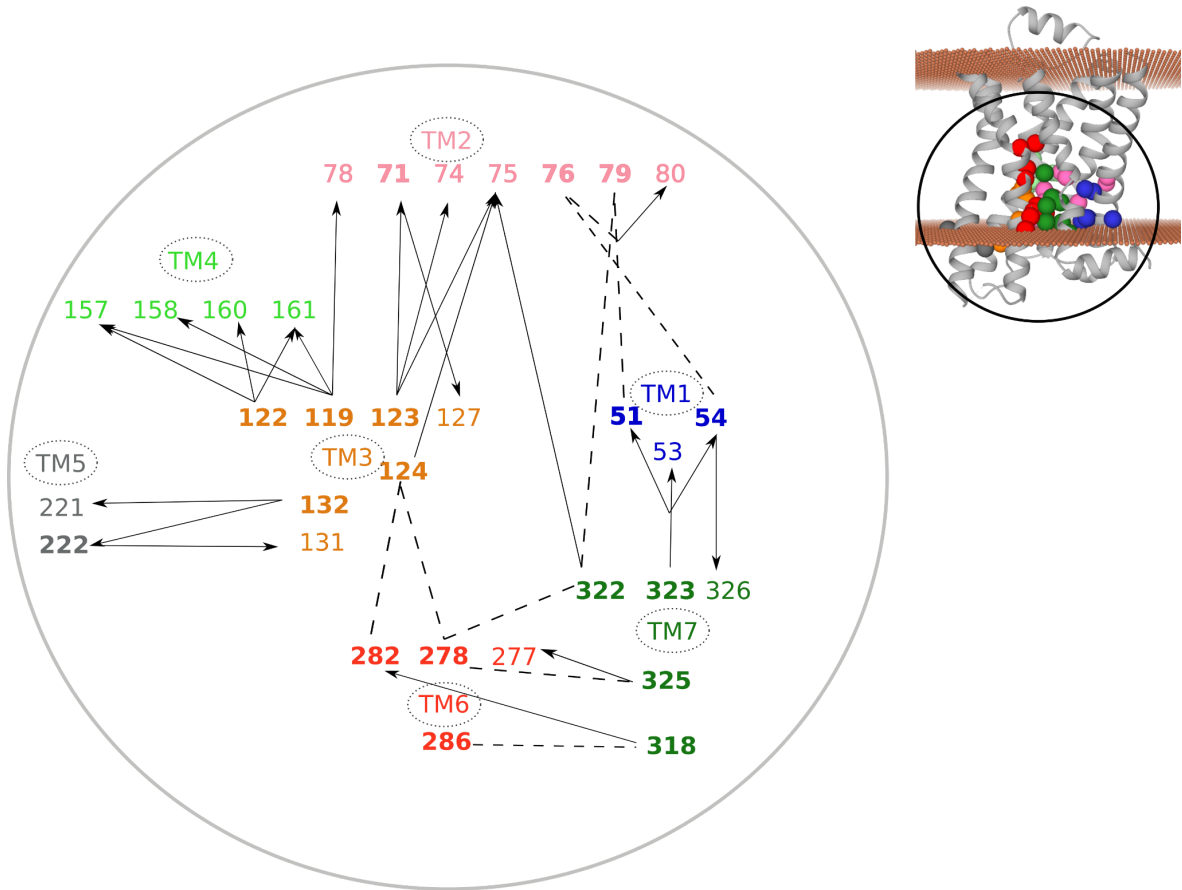


Figure 8: Top right corner shows A representative GPCR in gray ribbon with the planes of small orange spheres indicating membrane boundaries as predicted by the PPM server[82]. The spheres indicate the consensus conserved r groups from the inactive state analysis. The inset shows that the consensus conserved r groups are located primarily in the membrane embedded region. The magnified inset is shown as the large circle C that shows contacts between the residues of consensus conserved r-group from the inactive state analysis. Numbers are residue numbers of 2RH1 PDB. The six transmembrane helices labeled as TM 1 to 7 are represented as 6 different colors. The residues are coloured according to the <sup>TM</sup> helix they belong to. Numbers in bold indicate the central residue of the clique. Dotted lines indicate that the residues are conserved in each other's clique. For instance, residue number 124 from TM3 is a part of clique of residue number 282 from TM6 and vice versa. Solid black line indicates member of the clique. For instance, residue 75 from TM2 is a part of clique of residue number 124 from TM3

### 3.3.3 Active state analysis

Similar to the inactive state analysis, we analyzed 15 active state structures (Table 1) using three active state reference structures (Table 2) to identify regions of geometric and chemical similarities. Similar to the inactive state reference structures, the active state reference structures also were selected such that they had different resolutions and responded to a different type of ligand. We found that 22, 31 and 25 cliques were conserved for all 3 reference structure analysis respectively (Table 4). 19 cliques are common to at least 2 reference structures and 10 cliques are common to all 3 reference structures (Table 4). We found that 69 r-groups (Figure 6B) are conserved in at least 2 reference structure analysis and 37 are conserved in all the 3 reference structure analysis (Figure 9 - B). The 69 consensus conserved r-groups are contributed by 28 amino acids. The BW numbering of the 28 amino-acids that contribute to the consensus conserved r-groups is '1x50', '1x53', '1x54', '2x42', '2x45', '2x46', '2x47', '2x49', '2x50', '2x51', '3x39', '3x48', '3x49', '3x50', '3x51', '3x54', '4x50', '4x53', '5x47', '5x53', '5x58', '5x60', '5x61', '6x40', '6x48', '7x45', '7x49', and '8x50'.

Table 4: Consensus cliques conserved from the active state in reference structure1, reference structure2, reference structure3. The last column represents the frequency of consensus conserved cliques from the active state in the 3 reference structures. The star\_no column indicates the central residue of the clique. ‘Conserved\_in\_active\_reference structure1’ column indicates the conserved residue number and the conserved r-group separated by ‘\_’. All the residue numbers follow the numbering of PDB 2RH1.

star_no	conserved_in_active_reference1	conserved_in_active_reference2_equivalent_to_grid1	conserved_in_active_reference3_equivalent_to_reference1	How_many_active_references_it_is_conserved
55	['80_r8', '84_r2', '84_r8']	['80_r8', '83_r6', '84_r2', '84_r8', '299_r7']	['80_r8', '83_r6', '84_r2', '84_r8']	3
73	['70_r7', '76_r12', '76_r8']	NA	NA	1
80	['55_r2', '55_r9', '58_r8', '59_r8']	['55_r2', '55_r9', '58_r8', '59_r8']	['55_r2', '55_r9', '58_r8', '59_r12', '59_r8']	3
83	['55_r2', '55_r9', '124_r7']	['55_r2', '55_r9', '124_r7', '302_r2']	['55_r2', '55_r9', '124_r7', '299_r7', '302_r2']	3
87	['55_r2', '55_r9', '299_r7']	NA	NA	1
123	['78_r7', '82_r8', '164_r7']	['82_r8', '161_r2', '161_r15']	['82_r8', '161_r2', '161_r15', '164_r7']	3
124	['83_r6', '302_r2', '302_r9']	NA	NA	1
127	['75_r2', '78_r7', '79_r2', '79_r12', '79_r8', '82_r8']	['75_r2', '78_r7', '79_r2', '79_r12', '79_r8', '82_r8']	['78_r7', '79_r2', '79_r12', '79_r8', '82_r8']	3

132	['218_r12', '218_r8', '219_r12', '219_r8', '223_r2']	['218_r12', '218_r2', '218_r8', '219_r13', '223_r2']	NA	2
135	['139_r12', '139_r8', '139_r2', '226_r8']	NA	NA	1
136	['225_r2', '226_r8']	NA	['225_r2', '226_r8']	2
139	['135_r2', '226_r8']	NA	NA	1
140	['136_r2', '225_r2', '226_r8']	NA	['136_r2', '136_r16', '225_r2']	2
164	['122_r2', '126_r2', '168_r7']	NA	NA	1
187	['103_r2', '103_r15', '110_r10']	NA	NA	1
222	['133_r12', '133_r8', '135_r2', '136_r2', '136_r16']	['133_r12', '133_r8', '136_r2', '136_r16']	['132_r8', '133_r12', '133_r8', '136_r2', '136_r16']	3
226	['135_r2', '136_r2', '136_r16', '139_r12', '139_r8', '139_r2', '254_r8']	NA	['135_r2', '136_r2', '136_r16', '139_r8', '254_r8']	2
269	['212_r2', '212_r14', '265_r2', '265_r15']	['212_r2', '212_r14', '213_r2', '265_r2']	['212_r2', '212_r14', '265_r2']	3
302	['79_r2', '79_r12', '79_r8', '83_r6', '124_r7', '306_r2', '306_r16']	['79_r2', '79_r12', '79_r8', '83_r6', '306_r2', '306_r16']	['79_r2', '79_r12', '79_r8', '83_r6', '306_r2', '306_r16']	3
303	['55_r2', '55_r9', '58_r8', '80_r8', '83_r6']	['55_r2', '55_r9', '58_r12', '58_r8', '80_r8']	['55_r2', '55_r9', '58_r8', '80_r8', '83_r6']	3



306	['76_r12', '79_r2', '79_r12', '79_r8', '302_r2', '302_r9']	['76_r2', '79_r2', '79_r12', '79_r8', '302_r2', '302_r9']	['76_r12', '79_r2', '79_r12', '79_r8', '257_r8', '302_r2']	3
307	['58_r12', '58_r8', '76_r12', '313_r2']	['58_r8', '76_r2', '313_r2']	NA	2
59	NA	['80_r8', '84_r2', '84_r8']	NA	1
68	NA	['62_r12', '62_r8', '62_r2', '313_r2']	NA	1
72	NA	['134_r6', '135_r2']	['76_r2', '134_r6', '135_r2']	2
79	NA	['128_r8', '131_r8', '302_r2', '302_r9', '306_r2', '306_r16']	['302_r2', '306_r2', '306_r16']	2
84	NA	['55_r2', '55_r9', '59_r8']	NA	1
86	NA	['91_r8', '124_r7', '296_r2']	NA	1
110	NA	['103_r2', '103_r15', '105_r2', '187_r10']	NA	1
125	NA	['212_r2', '212_r14', '265_r2']	NA	1
129	NA	['218_r12', '218_r8', '219_r8']	NA	1
131	NA	['75_r2', '76_r12', '76_r8', '306_r2', '306_r16']	NA	1
148	NA	['137_r2', '137_r8', '152_r2', '153_r8']	NA	1
153	NA	['71_r8', '74_r2', '75_r2']	NA	1
212	NA	['125_r8', '216_r2', '216_r12', '265_r2', '265_r15']	NA	1

215	NA	['125_r8', '126_r2', '128_r8']	NA	1
265	NA	['212_r2', '212_r14', '298_r2', '298_r9']	['212_r2', '212_r14', '261_r2', '298_r2']	2
298	NA	['261_r2', '261_r14', '264_r10', '265_r2', '265_r15']	NA	1
299	NA	['55_r2', '55_r9', '83_r6', '124_r7']	NA	1
305	NA	['256_r8', '257_r8']	['257_r12', '257_r8', '301_r12']	2
317	NA	['57_r2', '57_r8', '313_r2']	NA	1
219	NA	NA	['132_r8', '257_r12', '257_r8', '261_r2', '261_r14']	1
223	NA	NA	['132_r8', '254_r2', '257_r12', '257_r8', '258_r12']	1
230	NA	NA	['139_r12', '139_r8', '226_r8']	1
254	NA	NA	['223_r2', '226_r8']	1
261	NA	NA	['219_r12', '219_r8', '265_r2', '265_r15']	1
264	NA	NA	['294_r2', '294_r8', '298_r2']	1
267	NA	NA	['290_r2', '294_r2', '294_r8']	1
294	NA	NA	['263_r12', '263_r8', '263_r2', '264_r10', '268_r2']	1

Of these 28 residues, 18 residues, '1x50', '1x53', '2x42', '2x45', '2x46', '2x47', '2x49', '2x50', '2x51', '3x50', '4x50', '4x53', '5x60', '5x61', '6x40', '6x48', '7x45', and '7x49' are common with the consensus conserved from the inactive state. Although the residues are common, the contacts that they make with other residues may not be identical in both the states. For instance '1x50' is a part of the star of residue '2x47' in both active and inactive states. But in the active state an additional contact with '1x54' is observed in this clique (Please refer to Results section 3.3.4, 3.3.5 and 3.3.6 for more discussion).

We also found that all of the consensus conserved r-groups occur in at least 70% of our dataset(Figure 7B)

Similar to the inactive state conservation, ~90% of the consensus conserved r-groups for the active state lie within the membrane embedded region of GPCRs, with some conservation in the intracellular region, and no conservation in the extracellular region, similar to the observations from the inactive state (Figure 9). The extracellular domains include 3 extracellular loops that connect the transmembrane helices 2 with 3, 4 with 5 and 6 with 7. The extracellular loops are flexible in nature. One of the important functions of the extracellular domains of GPCRs is to bind/respond to various stimuli. The types of stimuli include photons, small molecules, proteins, peptides, lipids, mechanical stress. To sense such a wide variety of stimuli, the extracellular domains of GPCRs have also evolved to be structurally and chemically diverse. Additionally, when stimulated by ligand binding, the ligand binding site (which is often located in the extracellular domain) may or may not undergo conformational changes during the event of binding or during the process of activation. Thus, we were not expecting any conservation in the extracellular region.

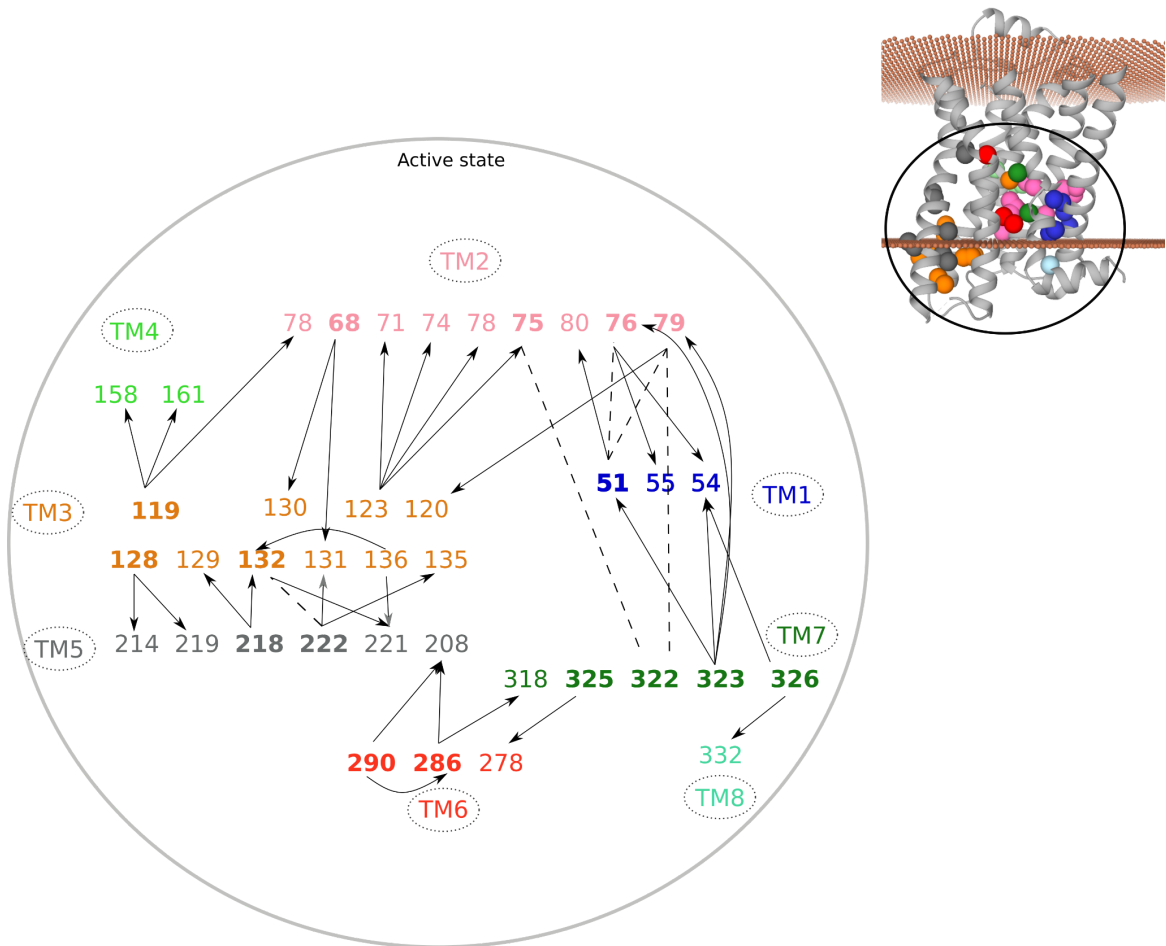
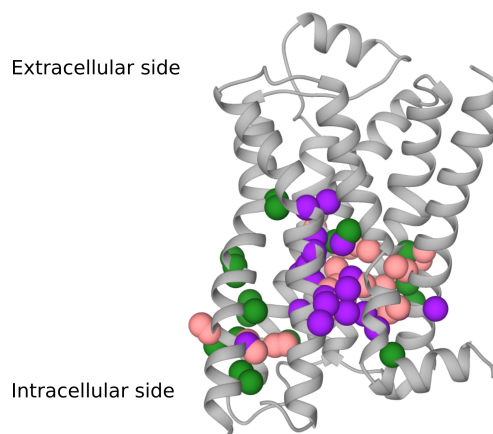


Figure 9: Top right corner shows A representative GPCR in gray ribbon with the planes of small orange spheres indicating membrane boundaries as predicted by the PPM server[82]. The spheres indicate the consensus conserved r groups from the inactive state analysis. The inset shows that the consensus conserved r groups are located primarily in the membrane embedded region. The magnified inset is shown as the large circle that shows contacts between the residues of consensus conserved r-group from the inactive state analysis. Numbers are residue numbers of 2RH1 PDB. The six transmembrane helices labeled as TM 1 to 7 are represented as 6 different colors. The residues are coloured according to the <sup>TM</sup> helix they belong to. Numbers in bold indicate the central residue of the clique. Dotted lines indicate that the residues are conserved in each other's clique. For instance, residue number 222 from TM5 is a part of clique of residue number 132 from TM3 and vice versa. Solid black line indicates member of the clique. For instance, residue 131 from TM3 is a part of clique of residue number 222 from TM5.

### 3.3.4 The r-groups that undergo conformational changes during activation

The cliques that are conserved in the inactive state but not in the active state are the ones that have undergone conformational changes during receptor activation. Such cliques can be considered important for activation. Out of the 18 cliques conserved in the inactive state analysis, 15 cliques are disrupted, either partially or completely (Table 5). These 15 cliques are centered at residues 1x53, 2x42, 2x50, 3x38, 3x41, 3x43, 3x51, 6x40, 6x44, 7x49, and 7x52. For each of these cliques, at least one consensus conserved r-group changed its location during activation. An example of a completely disrupted clique is a clique of residue 1x53 has 3 r-groups ('76\_r8', '326\_r2', '326\_r16') conserved in the inactive state analysis. In the active state analysis, this clique is not conserved, which indicates that this clique was completely disrupted during activation, where '76\_r8' relocated itself to a clique of 7x50, while 326\_r2 and 326\_r16 are not a part of any conserved clique in the active state. '2x47' clique is an example of a partially disrupted clique. In the inactive state, it contains 4 conserved r-groups ('51\_r2', '51\_r9', '54\_r12', '54\_r8'), of which only 1 r-group ('54\_r12') changed its location during activation, while the other 3 r-groups maintained their geometric positions. Finally, the 3 cliques that remain completely intact during activation have their stars at 1x50, 3x42, 5x61, indicating a possibility of their role in structural stability (Figure 10).



*Figure 10: A representative GPCR is represented in gray ribbon. The purple coloured spheres indicate the r-groups that are consensus conserved in the inactive state but not in the active state. The green colored spheres indicate the consensus conserved r-groups from the active state analysis but not in the inactive state. The pink coloured spheres represent the consensus conserved r-groups that are conserved in both the inactive state and the active state analysis.*

Table 5: Cliques that undergo conformational changes during activation. ‘Star\_no’ column indicates the central residue number, ‘inactive\_only’ indicates consensus conserved r-groups from the inactive analysis, ‘conserved\_in\_iaa\_not\_in\_aaa’ indicates the r-groups that are consensus conserved in the inactive state but not in the active state, ‘conserved\_aaa\_not\_in\_iaa’ indicates the r-groups that are consensus conserved in the active state but not in the inactive state, ‘conserved\_in\_aaa’ indicates the r-groups that are consensus conserved in the active state, ‘conserved\_in\_iaa\_and\_in\_aaa’ indicates the r-groups that are consensus conserved in both, the active state and the inactive state, ‘gpcr\_bw’ column indicates the BW numbering of the ‘star\_no’. All the residue numbers follow the numbering of PDB 2RH1.

star no	inactive only	conserved_in_iaa_not_in_aaa	conserved_aaa_not_in_iaa	conserved in aaa	conserved_in_iaa_and_in_aaa	gpcr bw
51	['76_r8', '79_r6', '80_r8']	NA	['80_r2']	['76_r8', '80_r2', '80_r8', '79_r6']	['76_r8', '79_r6', '80_r8']	1x50
54	['76_r8', '326_r2', '326_r16']	['76_r8', '326_r2', '326_r16']	NA	NA	NA	1x53
68	NA	NA	['130_r6', '131_r2']	['130_r6', '131_r2']	NA	2x39
71	['127_r12']	['127_r12']	NA	NA	NA	2x42
75	NA	NA	['322_r2']	['322_r2']	NA	2x46
76	['51_r2', '51_r9', '54_r12', '54_r8']	['54_r12']	['55_r8']	['51_r2', '51_r9', '54_r8', '55_r8']	['51_r2', '51_r9', '54_r8']	2x47
79	['51_r2', '51_r9', '322_r2', '322_r9']	['322_r9']	['120_r7']	['51_r2', '51_r9', '120_r7', '322_r2']	['51_r2', '51_r9', '322_r2']	2x50
119	['78_r8', '157_r8', '158_r2', '158_r15', '161_r7']	['157_r8']	NA	['78_r8', '161_r7', '158_r2', '158_r15']	['78_r8', '158_r2', '158_r15', '161_r7']	3x38
122	['157_r8', '160_r2', '161_r7']	['157_r8', '160_r2', '161_r7']	NA	NA	NA	3x41
123	['71_r2', '75_r2', '75_r12', '75_r8', '74_r7']	NA	['78_r8']	['71_r2', '74_r7', '75_r2', '75_r12', '75_r8', '78_r8']	['71_r2', '75_r2', '75_r12', '75_r8', '74_r7']	3x42
124	['75_r8', '278_r12', '278_r8', '282_r2', '282_r14']	['75_r8', '278_r12', '278_r8', '282_r2', '282_r14']	NA	NA	NA	3x43
128	NA	NA	['214_r12', '214_r8', '219_r2']	['214_r12', '214_r8', '219_r2']	NA	3x47
132	['221_r2', '222_r12']	['222_r12']	NA	['221_r2', '222_r8']	['221_r2', '222_r8']	3x51

	['222_r8']					
136	NA	NA	['132_r2', '221_r2']	['132_r2', '221_r2']	NA	3x55
218	NA	NA	['129_r12', '129_r8', '132_r2', '132_r16']	['129_r12', '129_r8', '132_r2', '132_r16']	NA	5x57
222	['131_r2']	NA	['132_r2', '132_r16', '135_r8']	['131_r2', '132_r2', '132_r16', '135_r8']	['131_r2']	5x61
278	['124_r8', '322_r2', '322_r9', '325_r8']	['124_r8', '322_r2', '322_r9', '325_r8']	NA	NA	NA	6x40
282	['124_r2', '124_r8']	['124_r2', '124_r8']	NA	NA	NA	6x44
286	['318_r9']	['318_r9']	['208_r2', '208_r14', '318_r2']	['208_r2', '208_r14', '318_r2']	NA	6x48
290	NA	NA	['208_r2', '208_r14', '286_r2']	['208_r2', '208_r14', '286_r2']	NA	6x52
318	['282_r2', '282_r14', '286_r2', '286_r15']	['282_r2', '282_r14', '286_r2', '286_r15']	NA	NA	NA	7x45
322	['75_r2', '75_r8', '79_r6', '278_r12', '278_r8']	['278_r12', '278_r8']	['75_r12']	['75_r2', '75_r12', '75_r8', '79_r6']	['75_r2', '75_r8', '79_r6']	7x49
323	['51_r2', '51_r9', '53_r2', '54_r12', '54_r8']	['53_r2', '54_r12']	['76_r8', '79_r6']	['51_r2', '51_r9', '54_r8', '76_r8', '79_r6']	['51_r2', '51_r9', '54_r8']	7x50
325	['277_r8', '278_r12', '278_r8']	['277_r8', '278_r12']	NA	['278_r8']	['278_r8']	7x52
326	NA	NA	['54_r8', '332_r2']	['54_r8', '332_r2']	NA	7x53

### 3.3.5 Newly formed contacts for stabilizing the active state

To check if new contacts are formed that assist stabilization of the receptors in their active state, we identified the consensus conserved cliques from the active state that are not conserved in the inactive state (Table 5). The active state has 18 cliques that are consensus conserved. These cliques are centered around 1x50, 2x39, 2x46, 2x47, 2x50, 3x38, 3x42, 3x47, 3x51, 3x55, 5x57, 5x61, 6x48, 6x52, 7x49, 7x50, 7x52, and 7x53 residues. Of these, seven cliques of residues 2x39, 2x46, 3x47, 3x55, 5x57, 6x52, and 7x53 are newly formed in the active state. 3x38, 3x51, and 7x52 are the cliques that do not form new contacts, instead they lose contacts during activation.

The remaining cliques gain contacts/neighbors during the process of activation(Figure 10). To summarize, contacts of 30 r-groups are newly formed in the active state(Table 5).

### 3.3.6 Cliques important for structural stability

The r-groups that are conserved in both the states, inactive and the active state, as a part of the same clique, indicate that they have not undergone structural changes during activation. We consider such r-groups to be important for maintaining the structural stability of the receptors. We found that 10 cliques maintain at least 1 and a maximum of 5 such r-groups that do not change their position(Table 5, Figure 11). These cliques are centered at residues with BW numbering as '1x50', '2x47', '2x50', '3x38', '3x42', '3x51', '5x61', '7x49', '7x50', and '7x52' (Figure 10). All these stars of the conserved cliques belong to transmembrane helix 1, 2, 3, 5 and 7. Interestingly, no clique from TM6 is seen to be conserved in both active and the inactive state, which is known to undergo conformational changes during activation. [86] have also reported TM6 to be conformationally flexible. This finding adds confidence to the results that we have obtained. Each of these cliques have at least 1 and a maximum of 5 r-groups that are conserved in both the states. Cliques that are centered around 7x52, and 5x61 have only 1 r-group that is conserved in both the states, indicating that the rest of the clique has undergone conformation changes. Further detailed analysis of such cliques is necessary to confirm their role in structural stability. If we only consider cliques that have at least 2 r-groups conserved, we notice that these r-groups belong to helices other than TM3 and TM6. Since, it has been previously shown that the contacts between TM3 and TM6 are lost during activation, these findings strengthen the results that we have obtained about the cliques important for structural stability and also increase confidence in the robustness of the algorithm that we have designed to obtain the conserved cliques.



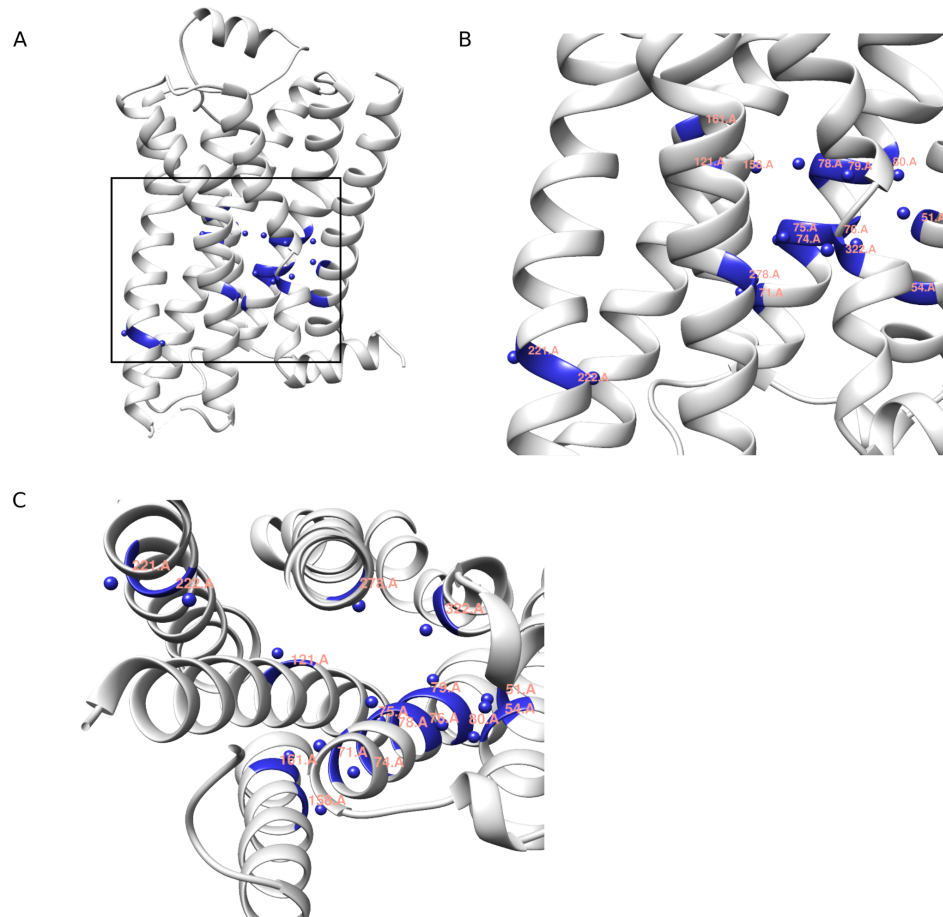


Figure 11: A)  $C^{\beta}$  of the residues that are important for structural stability are represented as small spheres rendered in navy blue color. The inset is magnified in (B). C) Bottom view indicating location of the structurally important residues.

### 3.3.7 Validation using data from literature

We used data from Zhou et al,2019[87] that performed 35 mutations on 17 residues of the Adenosine A2A receptor (A2AR). In this study, wildtype like ligand potency is considered as non-deleterious while a change in ligand potency of higher than 10 fold (increase or decrease) is considered as deleterious. The predicted phenotype for our study is based on the presence or the absence of the consensus conserved r-group in the mutated amino acid. For instance, if r7 is a consensus conserved r-group, its neutral mutation would be serine or threonine as these amino acids contain r7. Mutations to all other amino acids (except serine and threonine) would be

predicted as deleterious as they do not contain r7. We found that 11 of the 17 residues overlapped with the conserved residues in the inactive state from our study. These 11 residues contribute to 26 mutations. Of these 26, 16 mutations agreed with our predicted phenotype, while 10 did not (Table 6). The 10 mutations that did not agree with our predictions are contributed by 4 residues. Of these 4 residues, residue (3x50) R of the 'DRY' motif contributes to 3 incorrect predictions, residue 6x40 contributes to 4, 6x44 contributes to 2 incorrect predictions and 7x45 contributes to 1 incorrect prediction. Interestingly, these four residues are buried in the membrane and are near to the cytoplasmic side of the receptor. Similarly, we found an overlap of 5 residues from the active state with that of the 17 residues mutated in the Zhou et al, 2019 [85]. These 5 residues contribute to 7 mutations. Of these 7 mutations, 3 mutations agreed with our predicted phenotype. Of the 4 mutations that do not agree, 3 are contributed by residue (3x50) R of the 'DRY' motif (similar to the observation from the inactive state analysis) which is a conserved sequence motif on the intracellular side of the receptor (Table 6-B).

It may be possible that in these cases the ligand potency might be unaffected, but the effect of the ligand binding on initiating a response (efficacy) could be affected. Thus a more appropriate measure of assessing our algorithm and its prediction would be to use ligand efficacy instead of the ligand potency.

Table 6: Agreement between our predictions (A) from the inactive state, (B) active state and the 26 mutations that are experimentally validated. ‘Position’ column indicates the BW numbering of the mutated residue. ‘Mutation’ column specifies details of the mutation. ‘Experiment result’ describes the phenotype of the mutation as observed in the experiments. ‘Mutated\_aa\_rgrp\_composition\_definition’ column indicates the r-group composition of the mutated amino acid, ‘original\_aa\_rgrp\_composition\_definition’ indicates the r-group composition of the wildtype amino acid. ‘tkmsm\_conserved\_rgrps’ indicates the r-group conserved as found in this study. ‘Conserved\_rgrp\_present\_in\_mutated\_aminoacid’ indicates if the conserved r-group is present in the mutated amino acid, [] means that the conserved r-group is not present in mutated amino acid. ‘Predicted\_outcome’ column indicates our prediction of activity of the receptor upon mutation.

A)

Position	Mutation	Experiment result	mutated_aa_rgrp_composition_definition	original_aa_rgrp_composition_definition	tkmsm_conserved_rgrps	conserved_rgrp_present_in_mutated_aminoacid	predicted_outcome
3x46	I98N	Low expression	['r2', 'r9']	['r12', 'r2', 'r8', 'r8']	['r12']	[]	Deleterious
3x46	I98E	CIM, >20-fold decrease in EC50	['r2', 'r6']	['r12', 'r2', 'r8', 'r8']	['r12']	[]	Deleterious
3x50	R102H	Close to WT	['r4']	['r2', 'r2', 'r3']	['r2']	[]	Deleterious
6x40	I238Q	Close to WT	['r2', 'r2', 'r9']	['r12', 'r2', 'r8', 'r8']	['r8', 'r12']	[]	Deleterious
6x40	I238E	Close to WT	['r2', 'r6']	['r12', 'r2', 'r8', 'r8']	['r8', 'r12']	[]	Deleterious

6x40	I238A	Close to WT	['r8']	['r12', 'r2', 'r8', 'r8']	['r8', 'r12']	['r8']	Deleterious
7x45	N280S	Close to WT	['r7']	['r2', 'r9']	['r9']	[]	Deleterious
3x50	R102A	Close to WT	['r8']	['r2', 'r2', 'r3']	['r2']	[]	Deleterious
6x40	I238M	Close to WT	['r2', 'r13']	['r12', 'r2', 'r8', 'r8']	['r8', 'r12']	[]	Deleterious
6x44	F242T	Close to WT	['r7', 'r8']	['r2', 'r14']	['r2', 'r14']	[]	Deleterious
6x44	F242L	Close to WT	['r2', 'r12', 'r8', 'r8']	['r2', 'r14']	['r2', 'r14']	['r2']	Deleterious
6x44	F242A	7.5-fold increase	['r8']	['r2', 'r14']	['r2', 'r14']	[]	Deleterious
3x43	L95A	Constitutively active	['r8']	['r2', 'r12', 'r8', 'r8']	['r8', 'r2']	['r8']	Deleterious
3x43	L95R	Constitutively active	['r2', 'r2', 'r3']	['r2', 'r12', 'r8', 'r8']	['r8', 'r2']	['r2']	Deleterious
6x40	I238Y	Constitutively active	['r2', 'r16']	['r12', 'r2', 'r8', 'r8']	['r8', 'r12']	[]	Deleterious
2x50	D52A	Completely abolished	['r8']	['r6']	['r6']	[]	Deleterious
6x44	F242R	373.6-fold decrease	['r2', 'r2', 'r3']	['r2', 'r14']	['r2', 'r14']	['r2']	Deleterious
6x48	W246A	219.9-fold decrease	['r8']	['r2', 'r15']	['r2', 'r15']	[]	Deleterious
7x45	N280R	Completely abolished	['r2', 'r2', 'r3']	['r2', 'r9']	['r9']	[]	Deleterious

2x46	L48R	Completely abolished	['r2', 'r2', 'r3']	['r2', 'r12', 'r8', 'r8']	['r8', 'r2', 'r12']	['r2']	Deleterious
3x43	L95F	7.8-fold decrease	['r2', 'r14']	['r2', 'r12', 'r8', 'r8']	['r8', 'r2']	['r2']	Deleterious
7x49	N284A	Completely abolished	['r8']	['r2', 'r9']	['r2', 'r9']	[]	Deleterious
7x49	N284K	Completely abolished	['r2', 'r2', 'r2', 'r5']	['r2', 'r9']	['r2', 'r9']	['r2']	Deleterious
3x46	I98A	23.2-fold decrease	['r8']	['r12', 'r2', 'r8', 'r8']	['r12']	[]	Deleterious
7x53	Y288A	16.1-fold decrease	['r8']	['r2', 'r16']	['r2', 'r16']	[]	Deleterious
3x50	R102L	10.1-fold decrease	['r2', 'r12', 'r8', 'r8']	['r2', 'r2', 'r3']	['r2']	['r2']	Non-deleterious

B)

Position	Mutation	Experiment result	mutated_aa_rgrp_composition_definition	original_aa_rgrp_composition_definition	tkmsm_conserved_rgrps	conserved_rgrp_present_in_mutated_aminoacid	predicted_outcome
2x46	L48R	Completely abolished	['r2', 'r2', 'r3']	['r2', 'r12', 'r8', 'r8']	['r2', 'r12', 'r8']	['r2']	Deleterious
2x50	D52A	Completely	['r8']	['r6']	['r6']	[]	Deleterious

		abolished					
3x50	R102H	Close to WT	['r4']	['r2', 'r2', 'r3']	['r2']	[]	Deleterious
3x50	R102A	Close to WT	['r8']	['r2', 'r2', 'r3']	['r2']	[]	Deleterious
3x50	R102L	10.1-fold decrease	['r2', 'r12', 'r8', 'r8']	['r2', 'r2', 'r3']	['r2']	['r2']	Non-deleterious
3x51	Y103E	Close to WT	['r2', 'r6']	['r2', 'r16']	['r16', 'r2']	['r2']	Deleterious
6x48	W246A	219.9-fold decrease	['r8']	['r2', 'r15']	['r2']	[]	Deleterious

	Increased ligand potency
	Decreased ligand potency
	Did not agree with our predictions
	Not validated due to low expression levels

Apart from the analysis of 35 mutations, we also validated our results using another set of 435 disease associated mutations that were collected from the literature and were collated by Zhou et al, 2019[87]. These 435 mutations are reported for mutations in various Class A receptors. We found that 94(~22%) of these 435 (of which 88 are not embedded in the transmembrane region) disease associated mutations overlap with the residues that are found important in our analysis. We further identified which of these 94 mutations are contributed by the inactive state and which of them are contributed by the residues conserved in the active state. We found that 79(~20% of the total 435 mutations) are contributed by the residues conserved in the inactive state, of which 15 predictions were incorrect. In 11 of these 15 incorrect predictions the r2 and/or r8 r-group was found to be conserved. Next, we found that 43 of the 94 overlapping mutations(~20% of the total 435 mutations) are contributed by the residues conserved from the active state. We found that this set had 9 incorrect predictions, of which the r2 r-group was found to be conserved in 6 of them. (Please refer <https://docs.google.com/spreadsheets/d/1Hwsy1DmIEIa6g8f9BSKZa5i5bSaHdMRU5Ri8X0hnx3Y/edit?usp=sharing> to access this dataset). 28 mutations are common for the inactive and the active state conservations.

Similar to the 35 mutations set, we find incorrect predictions for residue 3x50 in the set of 435 disease associated mutations collected from the literature. There are 12 mutations reported for this position, of which we predict only 4 incorrectly. However, we do not have a single incorrect prediction for the other 3 positions (6x40, 6x44, and 7x45) that were incorrectly predicted in the set of 35 mutations.

One reason for the incorrect predictions could be the r-group definition used in this study. The definition currently implemented might not account for some of the interactions, resulting in loss of meaningful interactions. One could take consensus of multiple such definitions of r-groups and then check if the prediction accuracy improves.

To summarize, we have successfully attempted to validate the residues predicted as important for the structure and function of GPCR using the data from the literature.

### 3.3.8 Testing using Molecular dynamics simulations

We predicted residues in the GPCRs are important for their activation. To test these residues, we designed two types of mutations, disrupting and non disrupting for 4 such residues (Table 7). The disrupting mutation causes a GPCR to be either constitutively active or constitutively inactive, leading to a disruption of its wild type activity. The second type is where the mutation does not cause significant change in the activity of the GPCR (activity is similar to the wildtype) and is termed as non-disruptive type of mutation.

As a control mutation, we selected another residue, F282 that was previously reported to show a disrupting and non-disrupting type of phenotype when mutated to R and A respectively[45]. F282R shows a ~300 fold decrease in the potency. F282A shows 8 fold increase, and hence we considered it as wild type activity. We selected 4 other residues that are identified as important from our study and designed disrupting and non-disrupting mutations for them (Table7). For each of these designed mutations, we performed MD simulations. We intend to design mutations and run MD simulations for other residues that are not validated using the data from the literature as well.

We ran 200ns all atom simulations on GROMACS and took snapshots every 5ns. Since we wanted to quantify the effect of the mutation on its local environment, we used a novel metric called clique RMSD. Clique RMSD is a measure of the fluctuations caused by the mutation on residues that are in its clique (Refer section 3.2.2.c for definition of clique). The RMSD of a clique is calculated with respect to the starting structure and is averaged over all snapshots for triplicates of each mutation. We then compared the clique RMSDs of the mutants with that of the wildtype. The deleterious type of mutations are expected to have clique RMSDs substantially different (could be higher or lower) from that of the wildtype, while the non disrupting mutations are expected to have clique RMSDs similar to that of the wildtype. We performed a paired t-test to test the significance of the differences in the clique RMSD of the mutants and the wildtype. We think that using clique RMSD as an assessment metric is more meaningful than the traditionally used metric of RMSD (global) for trajectory analysis. This is because the starting structure of the receptor is stabilized by an antagonist (as per the crystal structure) and hence the effect of the mutation globally on the receptor may not be evident in a trajectory of 200ns. Additionally, we believe that the effect of the mutation on its immediate local environment would be more pronounced than that seen globally.



We observed that in the control type of simulation, the disrupting mutation has clique RMSD different than the wildtype. The non-disrupting mutation too had clique RMSD different than the wildtype. However, this non-disrupting mutation shows 8 fold increase in the activity of the receptor. The differences were also statistically significant as assessed by paired t-test. Our first set of designed mutations A76M(disrupting) and A76T(non disrupting) have clique RMSD less than the wildtype and mutation had clique RMSD similar to that of the wildtype as assessed by the paired t-test. In the second set, the disrupting mutation F208D had a significantly different clique RMSD profile than that of the wildtype, but the non-disrupting type of mutation for this residue F208V too had a clique RMSD profile significantly different from its wildtype (Figure 12). In the third set, the disrupting mutation V54A had lower clique RMSD than the wildtype while the non disrupting mutation V54I had clique RMSD higher. Finally, the fourth set, the disrupting type S161G showed a clique RMSD less than that of the wildtype while the non disrupting type S161T had higher clique RMSD than the wildtype.

Next, we calculated RMSF (root mean square fluctuation) of the mutated residue and compared it with the RMSF values of the wildtype residue (Figure 13). RMSF quantifies the movement of a residue (group of atoms) over a simulation trajectory. Higher RMSF values indicate higher fluctuations of the residue. In our study, we compared the RMSFs of the mutated residues with those of the wildtype to gauge the stability of the residue in its local environment. We hypothesized that residues that have RMSFs similar to the wildtype are stabilizing. Similarly, if the mutation is not stable in a given local environment, it will show RMSFs that are different from the wildtype. In our control mutation F282R/A, the disrupting mutation F282R showed higher RMSFs than the wildtype receptor, indicating that this mutation is not stable in its local environment. Its non-disrupting mutation, F282A, shows RMSF slightly less (0.02 units less) than the wildtype. In our first set of designed mutations, A76M/T, the disrupting and non-disrupting mutations both show similar RMSFs that are higher than the wildtype. This observation is consistent with its clique RMSD profile as well (Figure 12). This may indicate that the non-disrupting mutation is actually destabilizing the receptor. One of the reasons for this could be that the non-disrupting mutation A76T has an r7 group in addition to the conserved r8 r-group. This additional r7 group may be destabilizing the local environment. The second set of mutations, F208D/V, shows the RMSFs of disrupting mutations F208D higher than that of the wildtype. Its non-disrupting mutation, F208V shows RMSFs similar to the wildtype indicating that it is not adversely affecting the local environments. The third set of mutations, V54A/I, shows similar RMSF values for the wildtype and non-disrupting type of mutations. Its disrupting mutation, V54A, shows lower RMSF than the wild type, indicating that the local environment is not similar to that of the wildtype. In the final and fourth set of mutations that we

designed, S161G/T, the disrupting mutation S161G show higher RMSFs than the wildtype, and the non-disrupting mutation S161T shows RMSFs similar to that of the wildtype, indicating that the designed mutations are having the desired effect on the receptor.

To summarize, none of the designed mutations showed expected trends when we used clique RMSD metric to analyze the trajectory. When we used residue RMSF values, 3 of the 4 sets of designed mutations showed the desired trends, increasing confidence in our data. We believe that using multiple such assessment/analysis metrics could be useful in further validating our data.

Table 7: Mutations designed for testing the predictions validation using MD simulations

Sr. No	Residue Number	Wildtype ( $\beta$ 2AR)	Disrupting mutation	Non disrupting mutation
1	76	ALA (r8)	MET (r2, r13)	THR (r8)
2	208	PHE (r2, r14)	ASP (r6)	VAL (r12, r8) observed in a wild type receptor
3	54	VAL (r12)	ALA (r8)	ILE (r12)
4	161	SER (r7)	GLY (r1)	THR (r7)
5	282 (Control)	PHE (r2/r8, r14)	ARG (r3)	ALA (r8) - 8 fold increase observed in a wild type receptor

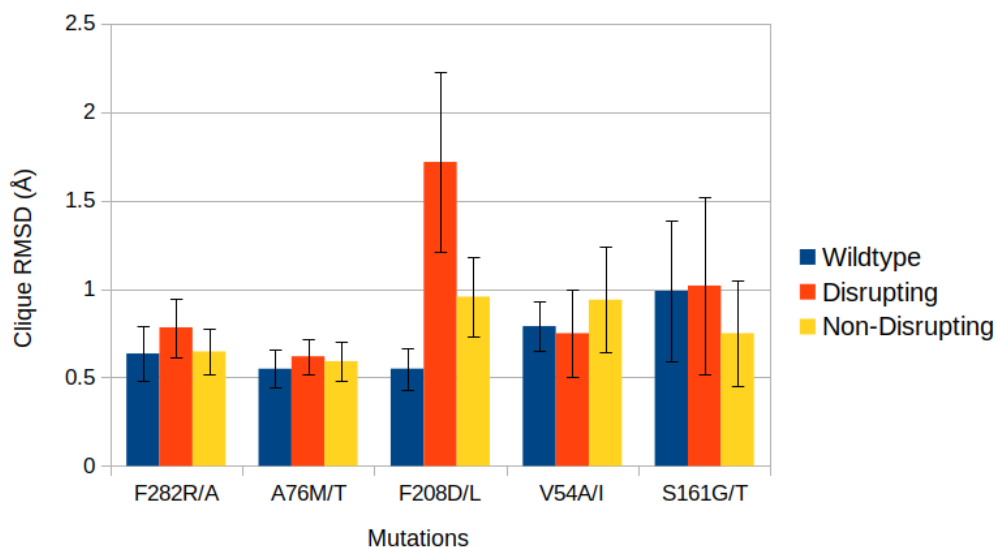


Figure 12: Clique RMSD values calculated using the MD simulation trajectories. The X axis indicates

mutation details in the format wildtype residue followed by disrupting mutation and non-disrupting mutation. For instance, F282R/A means F282 wildtype residue was mutated to R as a disrupting mutation and to A as a non-disrupting mutation.

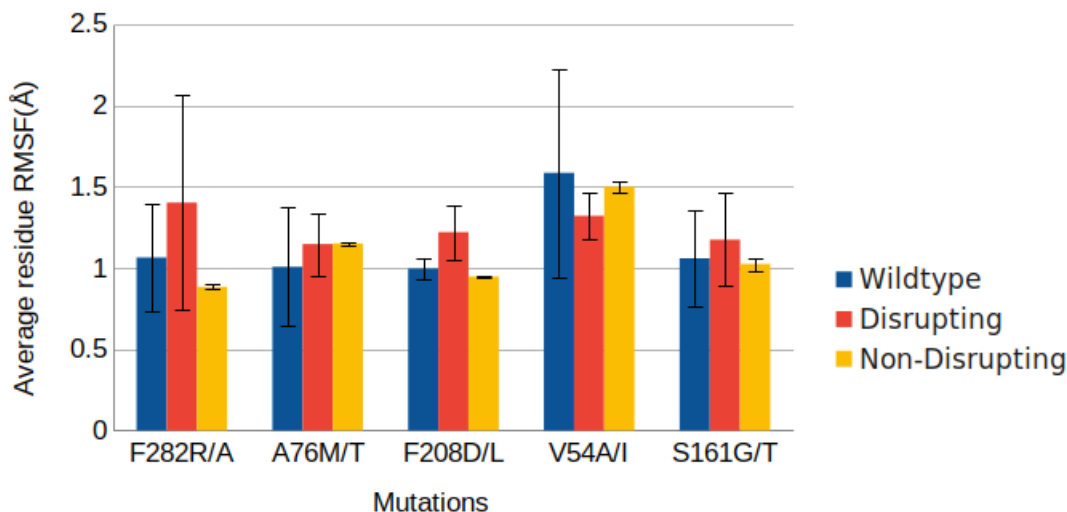


Figure 13: Residue RMSF values calculated using the MD simulation trajectories. The X axis indicates mutation details in the format wildtype residue followed by disrupting mutation and non-disrupting mutation. For instance, F282R/A means F282 wildtype residue was mutated to R as a disrupting mutation and to A as a non-disrupting mutation.

### 3.4 Discussion

In this study, we designed an algorithm to identify conserved regions that have similar geometric and chemical properties in a set of proteins represented as r-groups. The algorithm identifies geometric similarities by local superimposition. The chemical similarities are gauged using Shannon entropy. The use of r-group definition and Shannon entropy allowed us to find similarities in a protein family that has highly diverged sequences. We applied this algorithm to a class of proteins, GPCRs. We analyzed GPCR class A structures to address if all the GPCRs undergo activation in a universal way and if not, in how many different ways the activation happens. Using the algorithm, we identified regions of conserved geometric and chemical properties, in both the inactive and the active state structures. We compared the conserved regions

from the inactive and the active state and predicted the r-groups that are necessary for either structure, or function or both. We then validated these findings using data from literature. We designed mutations and performed MD simulations for the residues that were not validated through literature.

The conserved cliques that are identified in this study is presumably the entire activation pathway. Even though the GPCRs are highly diverse, we believe that these cliques are conserved in most GPCRs, if not all. Each of the GPCR might have its own switch that connects the ligand binding region to the activation pathway, but the pathway itself is conserved. Because of which we think that GPCRs follow a modular architecture. Using this information, we can design a GPCR that responds to a ligand of our interest. The design will involve 2 essential steps: the first one is to design a ligand binding pocket that binds to a ligand of our interest, and the second part is combining this pocket with the activation pathway that we identified in our study. This designing principle could open tremendous opportunities in the field of drug design.

Our findings could be influenced by the availability of a limited number of structures and also the resolution of the membrane embedded region of the protein. Small changes in the location of the amino acid side chains could lead to different results. Along with this, the definition and size of the clique can also affect the results. For instance, cliques can be defined in multiple ways using different cutoff distances, fixed number of neighbors etc. Similarly, r-groups can also be defined in various ways. In this study, we treated the r11 of proline as a backbone group and hence our list of conserved residues did not include proline residues. Newer definitions of r-groups could probably address this. In addition to this, we have used heuristic cut offs for Shannon entropy, match frequency to find conserved cliques. We tried to overcome this limitation by taking a consensus of analysis using 3 distinct reference structures.

We validated our findings by showing overlap of our data with that of the data reported in literature. We attempted to validate some of the important residues using MD simulations for the residues that did not have data in the literature. We used clique RMSD as a metric of measuring stability of the mutated clique. We analyzed 200 ns trajectories. The clique RMSD profiles did not show expected trends for validation. We also analyzed the trajectories with another metric, RMSF. With this metric, 3 of the 4 sets of designed mutations showed desired trends. We believe that analyzing these trajectories with various such metrics could assist in better understanding our findings. One such metric could be the distance between TM3 - TM6. This particular metric could be helpful to observe if the receptor is undergoing activation (the distance increases during

activation). We think that simulating the system for longer durations may also be helpful. Additionally using various liganded, unliganded versions of the receptors as starting points for MDs could also provide additional insights. Another additional thing would be to analyze the MD simulation trajectories that are deposited in publicly available databases like the GPCRmd[88].

Previous MD studies have shown water molecules that are conserved in the membrane embedded region of GPCRs to be important for establishing a network of polar contacts. These water molecules are shown to be conserved across diverse GPCRs. In our study, we have not analyzed conservation of water molecules, but it would be interesting to see if the waters interact with conserved cliques and have any structural or functional implication [89].

In this study, we asked the question if the GPCR activation pathway is universal, and if not, in how many different ways can GPCRs be activated. Initially, when we analyzed the GPCR structures by representing them as amino acids, we did not find conserved cliques across different GPCRs. The only conservations we got were from the already well known motifs (both sequence and structural motifs). This indicates that the amino acid cliques are diverse in terms of their composition. We then represented the GPCR structures as substructures of amino acids, the r-groups, and found conserved cliques in ~70% of the GPCRs. This indicates that small regions of amino acids and not the entire amino acid play a crucial role in their activation. Thus, we may need to analyze the structures using various definitions/representations of r-groups to robustly identify the non-conserved regions. It would be worthwhile to look at the non conserved regions / GPCRs to get further insights into their functional evolution.

To identify the activation pathway, we analyzed the cliques that are unique to the inactive state and those that are newly formed in the active state. Another interesting category of cliques could be the ones that change during the activation/transition and relax to their original conformation once the activation is complete. With our algorithm, in its current state, it could be challenging to track such cliques, since we have considered the system to be in 2 states - inactive and active. Thus, if the inactive state and active state of clique is identical, we consider the clique to be not changing its geometry/composition, even though it has temporarily undergone changes during the process of activation. It is possible to track/identify such cliques if we consider the system to be in 3 states - inactive, intermediate and active. The PDB does have representation for the intermediate state of some of the receptors and test cases using such structures could be attempted. However, the intermediate stage at which the structure was captured might play an important role in such analysis. To overcome this limitation, analyzing MD simulation trajectories

could be useful. However, the MD simulations that we have performed in this study, do not show state transitions to track such cliques. In general, observing state transitions for bulky systems like GPCRs require running simulations at scale of microseconds and hence may pose a challenge.

While this study was carried out using Class A receptors, the study could be extended for other classes of GPCRs. Our study was limited to Class A structures as the other classes had low representation in the PDB. With the recent developments in the field, models predicted by methods like AlphaFold2[90], RosettaFold[91] etc. could be either used to derive or validate the activation pathway. This algorithm can also be applied to a variety of other problems that require finding conserved regions in proteins.

## Chapter 4

# Packpred: Predicting the functional effect of missense mutations

- Binary classification of missense mutations as neutral or deleterious
- Trained on T4 lysozyme saturation mutagenesis dataset
- Tested on CcdB saturation mutagenesis dataset and the Missense3D dataset
- Performs better than 6 other state-of-the-art methods

The statistical potential and the design of this study was performed by Kuan Pern Tan and is a part of his thesis also.

Published - Tan KP\*, **Kanitkar TR\***, Kwoh CK, M.S.Madhusudhan. Packpred: Predicting the Functional Effect of Missense Mutations. *Front Mol Biosci.* 2021 Aug 20;8:646288. doi: 10.3389/fmolb.2021.646288. PMID: 34490344; PMCID: PMC8417552.

[\* equal contributions]

## 4.1 Introduction

Amino acid substitutions could affect protein stability, alter/impair its function and possibly lead to disease conditions[92]. Several such single amino acid substitutions in proteins, also called missense mutations, are implicated in diseases such as cystic fibrosis, diabetes, cancer etc.[93, 94]. Data from clinical studies as well as from large-scale projects such as the Human Genome Project[95], HapMap Project[96], Exome Sequencing Project and the 1000 Genomes Project[97] unearth such single amino acid mutations. It would be instrumental to have a fast and automated computational method to accurately predict the functional effect of these mutations.

Several computational methods predict the effect of missense mutations. The methods utilize sequence or structure information or a combination of the two. The sequence based methods rely on previously known protein sequences and their characterizations deposited in databases. For example, in the SIFT method[98], mutational effect prediction is made based on a customized position specific substitution matrix (PSSM), constructed with PSI-BLAST[37] and MOTIF finder[99] to identify conserved local sequence regions. A majority of structure-based methods are based on machine learning algorithms that exploit different features. For instance, I-mutant2.0[100] is a support vector machine based method trained on features such as pH, temperature and mutation type. AUTO-MUTE 2.0[101] constructs a statistical contact potential with Delaunay tessellation and trained their models with additional attributes such as ordered identities of amino acids, pH and temperature. PoPMuSiC-2.0[102] uses a linear combination of 26 different statistical energy functions in an artificial neural network architecture. M-CSM[103] utilizes a graph metric to summarize physicochemical interactions within a cut-off distance as pattern signatures and trained them with Gaussian process regression model. SDM [104, 105], which does not rely on machine learning, constructs an environment-specific amino acid substitution matrix based on observed substitutions in evolutionary time. DUET[106] is a meta-algorithm that consolidates the methods of mCSM and SDM. Missense3D[107] is another structure based method that uses seventeen structural properties to predict the effect of the mutation. Dynamut2.0[108] uses normal mode analysis and graph-based signatures. In addition to the sequence and structure based methods, methods that use both information(hybrid methods) also exist. One such hybrid method is Polyphen[109]. It uses a modified PSSM, data from the Pfam database and structural features such as accessible surface area and amino acid volume to make a prediction. All these methods are able to capture some aspects (but not all) of how the



mutation affects the protein structure and/or function, as indicated by their predictive accuracy. One common problem for all these methods is the prediction of a high number of false negatives, which affects their overall predictive accuracy. Hence, despite these various efforts and algorithms, the functional fate of point mutations remains a challenging problem.

A missense mutation could lead to functional instability by either disrupting its structure or by affecting its interaction interface and/or active sites without necessarily impacting its structure. A mutational effect predictor should hence take into account both the effect of mutation on overall structural stability and on its functional relevance. In this study, we describe Packpred, a method/algorithm that addresses both these aspects. For structural features, Packpred uses an environment-dependent multi-body statistical potential and a depth dependent substitution matrix, FADHM. We had previously established that FADHM scores are useful in predicting the effects of point mutations[110]. However, similar to other methods, FADHM also suffers from overpredicting false negatives. The multi-body statistical potential considers the observed/expected ratio of cliques of residues. The greater the value of the ratio, the more energetically stable is the packing of amino acids in the residue clique. We further categorized these residue cliques based on their residue depths. Residue depth[41, 42] measures the degree of burial and hence the solvation effect on amino acids. Depth has been shown to correlate well with structural stability and free energy change of cavity-creating mutations in globular proteins[39, 40]. Our depth based statistical potential hence assesses the effect of mutation on local packing stability. To capture the functional relevance of amino acids, we used residue position Shannon entropy from a multiple sequence alignment of homologs of the query sequence. By this, we exploit evolutionary information to quantify the degree of observed variation at the position of mutation. Usually, the lesser the variation the greater the functional importance of the residue.

## 4.2. Materials and methods

### 4.2.1. Data sets

#### 4.2.1.1 Statistical potential data set

A set of 3753 protein structures with resolution better than 2.5 Å obtained from the Protein data bank (PDB)[111] was used to construct the clique statistical potential. The structures in this set are non-redundant at 30% sequence identity. To account for atomic position fluctuations (protein

dynamics) while considering amino acid cliques, 10 homology models were built using Modeller9.11[78] with the native protein serving as both target and template in a self-alignment. The ‘refine very slow’ option was used to relax the structures with maximum atomic flexibility. The ‘refine very slow’ option indicates the degree of refinement of a protein model through MD annealing. The very slow option allows for better refinement/modeling flexibility as compared to other options ‘very fast, fast and slow’. Selecting this option allowed us to better sample the alternate locations of atoms in the existing structure dataset for building a robust statistical potential. These homology models along with the native structure (i.e. 11 structures for each protein) were then used to build the statistical potential.

#### 4.2.1.2 Saturation mutagenesis data sets

Saturation mutagenesis data sets of two proteins, T4-lysozyme[112] and Controller of cell division or death B (CcdB)[113] were used in this study. T4 Lysozyme is a 164 amino acid residue protein with our reference structure being PDB: 2LZM that was solved at a resolution of 1.7 Å[114]. Each position except the first was mutated to 13 other amino acids (A, C, E, F, G, H, K, L, P, Q, R, S, and T). After excluding key catalytic site residues (D10, E11, R145, R148P) the data set consists of 1966 mutations. CcdB contains 101 amino acids and acts as a cytotoxin. Its structure was solved at 1.4 Å resolution (PDB: 3VUB[115]). Each position in CcdB mutated to all other 19 amino acids. A final set of 1534 mutations was obtained after removal of active site residues (I24, I25, N95, F98, W99, G100, I101),

#### 4.2.1.3 Missense3D data set

The Missense3D data set consists of 4099 mutations from 606 proteins extracted from Humsavar[116], ClinVar[117], and ExAC[107, 118]. Humsavar lists all the annotated missense variants from humans reported in UniProt and SwissProtKB. ClinVar catalogs variations in humans and their associated phenotype. ExAC is an exome aggregation consortium that describes the aggregation and analysis of human exome. The analysis includes quantification of the pathogenicity of variants. The data set of 4099 mutations consists of 1965 disease-associated variants and 2134 neutral variants (not associated with any known disease, yet). Packpred parameters were trained on the T4-lysozyme data set and tested on the CcdB and Missense3D data sets.

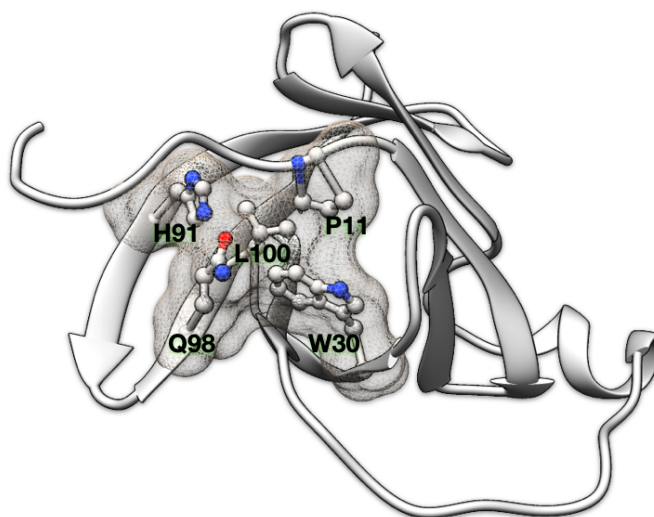
## 4.2.2 Structural and Sequential features

### 4.2.2.1 Residue depth

Depth is defined as the distance of a protein atom to the nearest bulk water molecule[41]. The quantity measures the degree of burial of the atom. Depth has been shown capable of concisely describing the protein environment, as substantiated by its utilities in protein design and function predictions[42, 110, 119]. Atom depth values were computed using default parameters. The depth of a residue clique is defined as the average depths of its constituent atoms.

### 4.2.2.2 Cliques of amino acid residues

A clique is defined as a sub-graph in which all possible pairs of vertices are linked. We define a  $(N, d_{\text{cut}})$  “residue clique” to be a clique of  $N$  amino acids within a linkage distance of  $d_{\text{cut}}$ . We consider two amino acids as linked when at least four, or more than half of side chain non-hydrogen atoms (whichever smaller) are within  $d_{\text{cut}}$  from atoms of another amino acid (Figure 1). For Glycine, the  $C^\alpha$  atom is used in lieu of the side chain. Residue cliques defined with different combinations of  $N$  and  $d_{\text{cut}}$  ( $N$  ranges from 2 to 4,  $d_{\text{cut}}$  ranges from 7.0 Å to 10.5 Å in step of 0.5 Å) have been computed and investigated in this study.



*Figure 1: Residue clique of amino acids. (A). A 5-residue clique (P11, W30, H91, Q98, L100) of cut-off 7.5Å shown in ball and stick representation and enveloped with a meshed molecular surface from human recombinant MTCP-1 protein (PDB: 1A1X).*

#### 4.2.2.3 Statistical potential and residue clique score

A residue clique statistical potential is constructed by adopting the formulation of Sippl's potential of mean force[120],

$$E^c = -kT \log \left( \frac{(P_{obs}^c + \alpha P_{exp}^c)}{(P_{exp}^c + \alpha P_{exp}^c)} \right)$$

(1)

Where  $E^c$  is the pseudo potential energy and  $c$  is a residue clique of type  $\{r_1, r_2, \dots\}$ , where the  $r_i$ 's are the amino acid types;  $P_{obs}^c$  is the observed number of residue clique  $c$ ;  $P_{exp}^c$  is its expected number in a hypothetical reference state without energetic interactions;  $\alpha$  is the ratio of pseudo-count introduced to account for sparse statistics, and is taken as 0.00 in our study.  $-kT$  is a constant and is assumed as 1 in this study.

For each  $(N, d_{cut})$  clique, the statistical potential is built at 5 different levels of depth (2.80 Å – 5.25 Å, 4.25 Å – 6.25 Å, 5.25 Å – 7.25 Å, 6.25 Å – 8.25 Å, 7.25 Å - ∞). To calculate the score of a residue clique (S), the mean  $\mu$  and standard deviation  $\sigma$  of its depth is first computed. A Gaussian probability density function  $N(x | \mu, \sigma)$  is then accordingly built. The clique score is computed as the weighted sum of the integrand at every depth level as,

$$S_{\mu, \sigma}^c = \sum_{d \in D} \frac{1}{d_f - d_i} \int_{x=d_i}^{x=d_f} E_d^c \cdot N(x | \mu, \sigma) dx$$

(2)

Where  $d$  is one depth level,  $d_i$  and  $d_f$  is the lower and upper bound of the level.

Most residue cliques in a protein are overlapping with one another, and an amino acid residue can participate in multiple cliques. The score of a residue is taken as the average of all such cliques. The score of a protein is further taken as the average of all its residue scores.

#### 4.2.2.4 Shannon Entropy

Shannon entropy (H) is a measure of variation observed at a given position. It is calculated from a multiple sequence alignment obtained by a PSI-BLAST search against the uniref50 database[37]. H for a given position is then calculated as,

$$H = -\sum_{i=1}^{20} P_i \log_2 P_i$$

(3)

Where  $P_i$  is the fraction of amino acid (i) observed at a given position.

#### 4.2.2.5 FADHM scores

FADHM scores are depth dependent pairwise amino acid substitution likelihood scores extracted from the FADHM matrices. The FADHM matrices quantify the substitution frequencies at 3 depth regions obtained by performing protein-protein structural alignments. The three depth regions are defined according to their residue depth values. The regions are called exposed, intermediate and buried. The exposed region has residues that have residue depth of less than 5, the intermediate region is characterized by depth value between 5 to 8 while the buried region has depth values higher than 8. The idea of creating such matrices is that the relative abundance of amino acids is different at different depths. This also implies that the substitution rates would also be different at different depths. The FADHM matrices were benchmarked using the saturation mutagenesis datasets of T4 lysozyme and CcdB. A detailed account of the FADHM score can be found elsewhere[110].

#### 4.2.2.6 The Packpred score for mutations

The Packpred score is given as,

$$PS = 1.5(S) + 1.75(H) + 0.5(FADHM) \quad (4)$$

Where PS is Packpred score, S (section 4.2.2.3) is the residue clique score obtained from the statistical potential, H is Shannon entropy (section 4.2.2.4) and FADHM (section 4.2.2.5) is the depth based amino acid substitution likelihood score. The weights were obtained by training on the T4 saturation mutagenesis data set. The coefficients for S, H and FADHM (weights) were systematically sampled in the range 0 to 3 with a step size of 0.25. The cut off score threshold that best discriminates neutral mutations from destabilizing ones was 1.6 in the training data. Mutation with a score greater than 1.6 is neutral and is destabilizing otherwise. To score a mutant, we modify the clique composition without explicitly modeling the mutant protein structure, with the mutant amino acid inheriting all the properties of the wild type residue.

Packpred is implemented as a web server at <http://cospi.iiserpune.ac.in/packpred/>. A standalone version is also available for download.

## 4.3 Results

### 4.3.1 Training and testing Packpred score

Packpred uses a linear combination of sequence position Shannon entropy, a residue clique statistical potential and a depth dependent substitution matrix (FADHM) to predict the functional effect of missense mutations. The Shannon entropy part of the score estimates the structural and functional importance of residues based on evolutionary information. The clique statistical potential and the substitution matrix gauge the effect of the mutation on the local environment/structure. The statistical potential computes the observed and expected probabilities to calculate a score for a clique. The FADHM scores are taken from substitution matrices that are derived from structural alignments of proteins. The substitution likelihood scores are calculated by categorizing a protein into three regions based on residue depths (exposed intermediate and buried). The substitution scores indicate the likelihood of a residue getting replaced by another at a given depth.

We performed a grid search in the range of 0 to 3 with a step size of 0.25 for S, H and FADHM to optimize the coefficients (weights) of each component of the linear combination Packpred score. The optimization was to maximize the Matthews correlation coefficient (MCC) (see section 2.3) T4 lysozyme saturation mutagenesis data training set. The weights that gave the highest MCC on the training set were 1.5, 1.75, and 0.5 for the clique statistical potential, Shannon entropy and FADHM respectively. We also obtained a cut-off threshold that distinguishes the destabilizing from the neutral ones from this training exercise. The cut off was sampled in the range 0 to 2 with a step size of 0.1. Mutations with scores greater than 1.6 are classified as neutral and scores below 1.6 are classified as destabilizing. The T4-lysozyme training set consists of 1362 (~69%) neutral and 604 (31%) destabilizing mutations of which Packpred correctly identifies 1049(~77%) neutral mutations and 406(~67%) destabilizing mutations . In the T4 training exercise, we observe similar MCC values for different combinations of weights of the grid search. Although the MCCs are similar, the underlying predictions and the linear combination scores are different.

The weights and threshold obtained from the training set were applied to two testing sets, CcdB

saturation mutagenesis data and Missense3D data set. CcdB data set has 1258 (~80%) neutral mutations and 276 (~20%) destabilizing while the Missense3D data set has 2134(~52%) neutral and 1965(~48%) disease mutations respectively. We used the PDB structures, 2LZM and 3VUB to obtain Packpred scores of T4-lysozyme and CcdB respectively. The biological unit of CcdB is a dimer and we did all the calculations using this dimeric state structure for CcdB. Packpred correctly predicts 864/1258(~68%) neutral and 253/276(~92%) destabilizing mutations from CcdB testing set and 1670/2134 (~78%) neutral, 1123/1965 (~57%) disease causing mutations from the missense3D data set.

We compared Packpred’s binary classification with several popular methods such as i-mutant2[100], mCSM[103], SDM[105], dynamut2[108], FADHM[110], and Missense3D[107] (Table 1). All the predictions were made using default parameters. Packpred was the best performing method on the T4-lysozyme training set and the Missense3D testing set with MCC values of 0.42 and 0.36 respectively. The next best method is Missense3D with MCC values of 0.40 and 0.33 for the T4 and Missense3D data sets respectively. The MCC of Packpred on the CcdB data set is 0.47 and is marginally outperformed by the best performing method, FADHM, which has an MCC of 0.48 (Table 1).

Table 1: Performance of some methods on T4, CcdB saturation mutagenesis and missense3D data sets. \*: Values taken from FADHM paper.

Method	MCC for T4 lysozyme saturation mutagenesis data set	MCC for CcdB saturation mutagenesis data set	MCC for Missense 3D data set
i-mutant 2.0	0.30*	0.36*	0.06
mCSM	0.22*	0.39*	0.05
SDM2	0.24*	0.33*	0.14
Dynamut2	0.09	0.15	0.06
Missense3D	0.40	0.39	0.33
FADHM	0.38*	<b>0.48*</b>	0.27
Packpred	<b>0.42</b>	0.47	<b>0.36</b>

The clique potential and FADHM were earlier trained on 3754 and 2384 PDB entries respectively. 89 of these PDBs are common to the 606 PDB entries that comprise the Missense3D testing set. These 89 overlapping entries include not just those that are identical but also those that are homologs (with sequence identities of 30% or greater). The overlapping PDBs account for 463 of 4099 mutations in the Missense3D dataset. Omitting these 463 mutations and using the

other 3636 mutations resulted in an MCC of  $\sim 0.37$ , comparable to the value of 0.36 obtained over the entire Missense3D data set of 4099 mutations.

#### 4.3.2 Analysis of the predictions on the Missense3D data set

The missense3D data set has a balanced representation of  $\sim 48\%$  disease-associated mutations and  $\sim 52\%$  neutral mutations. The data set however is skewed in terms of amino-acid abundance when compared to natural abundance. For instance, Arginine has the highest representation and accounts for  $\sim 16\%$  (664/4099) of the missense3D data set while its natural abundance is  $\sim 5\%$ . The next most abundant amino-acid in the Missense3D data set is glycine that accounts for  $\sim 9\%$  (372/4099) of the data (natural abundance is  $\sim 7\%$ ). The most frequent mutant is also Arginine (347/4099) followed by serine (343/4099). There are 2233 mutations in the exposed environment (depth less than 5 Å), 1258 in the intermediate environment (depth between 5 and 8 Å) and 608 in the buried environment (depth greater than 8 Å).

We assessed the performance of various methods on the Missense3D data set using metrics including sensitivity, specificity, precision, accuracy and f1 (Table 2). Packpred outperforms all other methods in MCC, precision and accuracy. Missense3D has the highest sensitivity and f1. Packpred has less sensitivity than FADHM and Missense3D indicating a scope of improvement. Packpred has a specificity of 0.57, indicating a higher number of false positive predictions. mCSM and i-mutant outperform all other methods in specificity. However, mCSM, i-mutant, SDM and dynamute predict a large number of false negatives (Table 3) that affects their MCC. Hence, we compare Packpred with FADHM and Missense3D in the next sections unless otherwise stated. Packpred has less number of false positives among FADHM, Missense3D and has the highest number of false negatives. The high false positive rate contributes to its lower specificity.



Table 2: The prediction performance of seven methods on the Missense3D data set. The best score in each assessment metric is shown in bold font.

Metric	Packpred	FADHM	Missense3D	Dynamut2.0	mCSM	i-mutant	SDM
MCC	<b>0.36</b>	0.27	0.33	0.06	0.05	0.06	0.14
Sensitivity (Class 0)	0.57	0.39	0.40	0.84	<b>0.92</b>	<b>0.92</b>	0.80
Specificity (Class 0)	0.78	0.85	<b>0.89</b>	0.20	0.10	0.12	0.34
Precision (Class 0)	0.71	0.71	<b>0.76</b>	0.49	0.49	0.49	0.52
F1(Class 0)	0.63	0.50	0.53	0.62	<b>0.64</b>	<b>0.64</b>	0.63
Sensitivity (Class 1)	0.78	0.85	<b>0.89</b>	0.20	0.10	0.12	0.34
Specificity (Class 1)	0.57	0.39	0.40	0.84	<b>0.92</b>	<b>0.92</b>	0.80
Precision (Class 1)	0.66	0.60	0.62	0.59	0.59	0.60	0.63
F1 (Class 1)	0.72	0.70	0.73	0.31	0.18	0.20	0.44
Accuracy	<b>0.68</b>	0.62	0.65	0.51	0.50	0.50	0.55

We analyzed the results structure-wise. Packpred correctly predicted all mutations from 264 (out of 606) structures and at least 50% mutations correctly from 507 structures. It could not correctly predict any mutation from 56 structures. In these 56 PDBs, the maximum mutations in any one protein were 4 while the average number of mutations per PDB is ~6. These 56 structures did not follow any particular discernible pattern or trait.

Packpred has limitations in several areas. One of which is its high number of false positive predictions which also affects its specificity. Other methods have a higher specificity but underperform in their sensitivity by overpredicting True Negatives. Packpred has fewer true positives as compared to FADHM and Missense3D, indicating another potential area for improvement. With more true positives, it is likely that Packpred's f1 value would also improve, which is currently bested by Missense3D. Packpred scored higher than 0.65 in all other metrics

(accuracy, precision, sensitivity and f1) indicating its overall balanced performance. We also calculated MCC for each native amino-acid type from the Missense3D dataset. We found that of all the 20 types of amino-acids, Packpred has the highest MCC of 0.40 for I, L and V amino-acids and lowest MCC for C with MCC of 0.17. Similar to Packpred, FADHM also has the lowest MCC of 0.04 for C amongst all the amino-acid types. FADHM has a best MCC of 0.47 for I, which also happens to be the single best MCC for an amino acid among other methods. Missense3D, in contrast to Packpred and FADHM, has the best prediction for C with MCC of 0.43 and has lowest MCC of 0.02 for W among other amino-acid types. These results show us amino acid wise prediction performance and thus giving scope for improvement

Table 3: Confusion matrix values for the different prediction methods. The values in bold font show the best in each category. TP, FP, TN and FN stand for True Positive, False Positive, True Negative and False Negative respectively.

Metric	Packpred	FADHM	Missense 3D	Dynamut2.0	mCSM	i-mutant	SDM
TP	1670	1816	<b>1890</b>	440	229	251	713
FP	842	1203	1177	312	<b>158</b>	164	420
TN	1123	762	788	1650	<b>1804</b>	1798	1542
FN	464	318	<b>244</b>	1685	1896	1874	1412

We stratified the missense3D data to particular depth zones (residues with depth values less than 5 are exposed, between 5 to 8 are intermediate and greater than 8 are buried) to assess the performance of these methods at particular depths. Packpred has 597/2233 (~72%) correct predictions from the exposed environment, 796/1258 (~63%) from the intermediate and 400/608 (~66%) from the buried environment. Packpred is the least accurate in predicting the effect of mutations in the intermediate environment. Interestingly, Missense3D is also the least accurate in this intermediate zone (Figure 2).

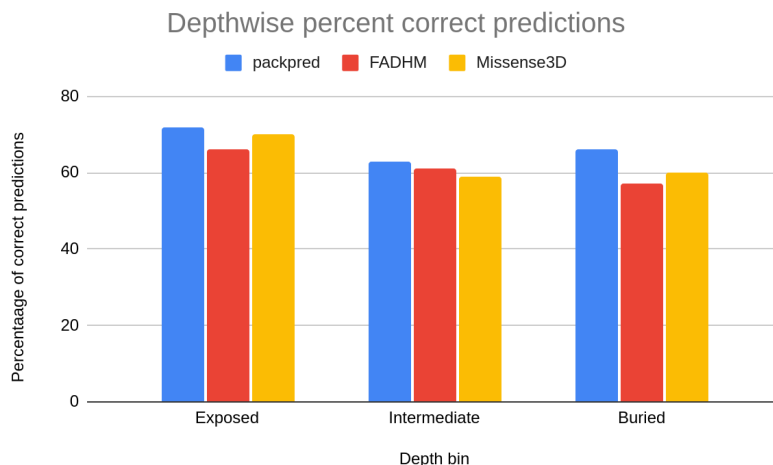


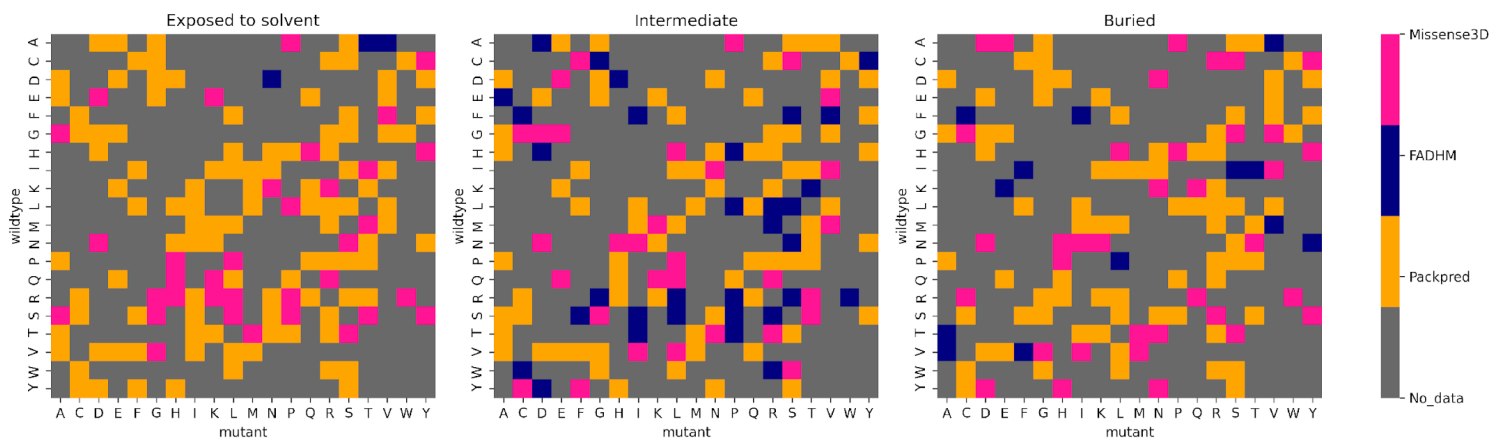
Figure 2: Histograms of the prediction accuracy of Packpred, FADHM and Missense3D at different depth levels (exposed to the solvent, intermediate and buried).

### 4.3.3 Meta predictions

Of the 4099 mutants, at least one of the seven methods we tested made an accurate prediction in 4036 cases. This motivated us to make two different meta predictions by combining the different methods.

The first meta prediction makes use of the method that performs the best for particular amino acids. We studied the wild type(native) amino acid-wise trends of all seven methods. For instance, native amino-acids N, K, Q, R and T are best predicted by Missense3D and FADHM outperforms other methods in the prediction of I and M amino acids and Packpred is best at predicting A, D, E, G, L, P, V, and Y. All seven methods feature as the best method for at least one amino acid. Interestingly, we found that Packpred has the highest average percentage (68%) of correct predictions of the 20 native amino acids with the lowest standard deviation (4%). In contrast, FADHM and Missense3D have averages of 62% and 64% with standard deviations of 7 and 10 respectively. The other methods all have averages less than 60% with standard deviations between 11-14. Packpred shows consistency in prediction across native amino acid types. We then used these prediction strengths of each of the methods to get a hypothetical hybrid/meta prediction scheme that combines predictions from all of the methods and has an MCC of 0.40 over the Missense3D data set), easily outperforming all the individual methods.

The second hypothetical meta prediction only involves Packpred, FADHM and Missense3D as these were the methods that did consistently well over all different data sets and amino acids. Here we considered the method that best predicted wild type-mutant pairs. Further, we segregated these amino acid pairs into different depth categories - exposed to solvent (depth < 5 Å), intermediate (depth between 5 to 8 Å) and buried (depth > 8 Å). Our meta prediction then chose the best performing method for a particular pair at a particular depth level. For instance, the wild type-mutant pair A→D, Packpred has the best predictions in an exposed environment, FADHM in the intermediate environment and Missense3D in the buried environment (Figure 3). In case of a tie between methods, the one with the better MCC was chosen. By thus combining the strengths of the three methods the MCC of the predictions rises to 0.51 for the Missense3D data set. An analysis to rationalize/explain why certain methods are best for certain pairs/environments did not yield any illuminating results. It is clear however that there is some degree of complementarity in these different methods and perhaps a more rigorous treatment of the results from the individual methods could further improve prediction accuracy.



*Figure 3: Best performing methods for each wild type-mutant amino acid pair at different depth levels.*

We would like to emphasize here that the purpose of exploring these meta predictions was to simply test the extent to which we could possibly improve results with such an approach. In a more rigorous implementation of this method we would have to train and test the meta-predictor separately, something that is beyond the scope of this study. Choosing the best results from our

testing set, as we have done here, merely represents the possible limit to which we could improve on predictions.

#### 4.3.4 Rank ordering the degree of phenotypic change by mutations

We wanted to investigate if the Packpred scores are indicative of the degree of change/disruption caused by a mutation. The degree of change is measured experimentally by the mutational sensitivity score, which categorizes each mutation into one of 4 and 8 levels in T4-lysozyme and CcdB data sets respectively. We chose to use Spearman's rank correlation coefficient (SCC) to measure the performance of rank-ordering, as it makes no assumption on linear relationship between the scores and the phenotypical change. SCC is calculated as

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

(6)

where  $d$  is the difference between the actual and the predicted ranks of a mutation, and  $n$  is the number of levels. The SCC for T4 and CcdB data sets is -0.48 and -0.54 respectively. At best, this correlation is weak and indicates that these scores could be further improved.

#### 4.3.5 Assessing robustness of Packpred

Lastly, we assessed the robustness of Packpred. For this, we changed the training set to include only 149 point mutations that result for a single nucleotide change in codons. The Missense3D dataset is made of only these 149 different mutations. We created three additional training sets that all contain instances of only these 149 mutations. The first contains mutations from only the T4 lysozyme dataset, the second set contains from T4 in a 50:50 ratio of neutral:deleterious mutations, and the third set has mutations from the T4 and CcdB datasets in the ratio 50:50 of neutral:deleterious. The ratio was chosen based on the neutral:deleterious ratio of the Missense3D test set. For every combination of the training set, we obtained different optimal weights for the features of the linear combination. Interestingly, the accuracy of the method as gauged by the MCC value over the Missense 3D dataset was consistently between 0.34 – 0.35.

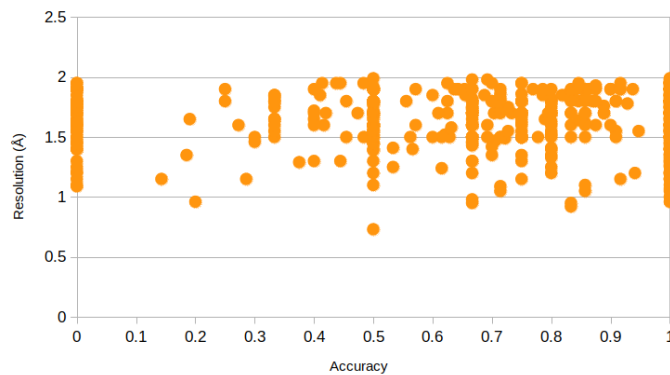
## 4.4 Discussions

In this study, we have developed a method to predict the effect of missense mutations on the structure and function of a protein. We believe that such predictions could be tested by assaying the protein for its function. Our method, Packpred, is constructed in a way that it is sensitive to structural changes effected by the mutation as well as any functional changes it may effect without perturbing the structures. To assess the impact of the mutation on the structure (and hence the function) of the protein, we devised a multi-body clique statistical potential. This statistical potential evaluates the strength of the interaction in a local neighborhood (amino acid clique). To assess the impact of mutation, we consider the same residue neighborhood environment while replacing the wild type amino acid with the mutant. The score of the clique with the wild type residue and with the mutant is computed. An inferior score for the mutant in comparison to the wild type would be indicative of a destabilizing mutation. The structural stability of introducing the mutant residue is also gauged by a depth dependent substitution matrix, FADHM, whose efficacy at detecting the fate of mutations we had previously benchmarked and tested. To account for functional changes caused by the mutation, we invoke evolutionary information from a multiple sequence alignment using Shannon entropy. The more conserved the position, the more likely it is going to affect function. These different scores are taken together in a linear combination, whose coefficients were optimized using the T4-lysozyme saturation mutagenesis data set of ~2,000 mutation. Packpred was tested on two different data sets, another saturation mutagenesis data set (CcdB) and the Missense3D data set. Its performance on these datasets was also compared to six other methods including FADHM, Missense3D, Dynamut2.0, mCSM, i-mutant2.0 and SDM. With an exception of the CcdB data set where it marginally underperforms FADHM, Packpred clearly outperformed all other methods on all data sets. Among the methods, Packpred balances well between predicting true positives and true negatives (neutral and disease causing mutations) and hence has the best MCC values. Packpred has the best accuracy and is close to the best specificity, precision and F1. It loses out to the best methods in these measures as well as on sensitivity as methods such as mCSM predict a disproportionately large number of negatives. When the performance of the different methods is compared on a (wild type) amino acid by amino acid basis, Packpred performs consistently well, with prediction accuracies never falling below 60% while maintaining an average of 68%, which is easily the best among the methods tested. Qualitatively, a similar picture also emerges when the results are broken down

into wild type-mutant amino acid pairs.

We also investigated whether Packpred (and other methods) preferred certain types of structures over others. No clear deduction could be made from these analyses. However, there was one trend that could be considered for further improvements – Packpred, similar to Missense3D and FADHM performed the worst in the intermediate amino acid depth environment. Mutational effects in exposed and buried (according to residue depth) environments were better predicted. Perhaps, the intermediate depth levels need to be further stratified, which in the case of Packpred would be reflected in the FADHM matrix values as well as in the clique statistical potential. There is scope of improvements for cases where Packpred was unable to accurately predict the fate of 72 mutants that were all accurately called by the other six methods. We could also dissect the 23 correct predictions that Packpred made that were missed by all other methods to determine the relative strength of Packpred in comparison to the other methods.

Packpred relies on the sequence and structure of a given protein to predict the effect of a mutation. These predictions could likely be impacted by the accuracy/resolution of the protein structure. The two structural features that Packpred extracts from structures are amino acid depth and structural neighbors. To whatever extent these two features get affected by the quality/accuracy/resolution of the structure would predicate the impact it would have on the final predictions. For the structures in the Missense3D dataset, they all have resolutions of 2 Å or better. For this set there appears to be no correlation between the accuracy of prediction and resolution of the structure (Figure4). In an independent study we are exploring the use of homology models along with low resolution structures from the PDB to quantify the impact of structural accuracy on Packpred predictions. We are also planning to run short MD simulations to extract snapshots of different conformations and use these snapshots for predictions. Such an exercise would help in better predictions for proteins that are highly flexible, such as the p53 protein.



*Figure4: Correlation between PDB structure resolution( $\text{\AA}$ ) and the accuracy of predictions by Packpred*

Packpred has limitations in several areas, one of which is its high number of false positive predictions which also affects its specificity. Other methods have higher specificity but underperform in the sensitivity by over predicting true negatives. Packpred also has less true positives as compared to FADHM and Missense3D data set indicating potential for improvement. Apart from MCC and sensitivity, Packpred has scores higher than 0.65 in all other metrics indicating its overall balanced performance. To assess which native amino acids predictions could be improved, we calculated MCC based on the prediction of native amino acids. We also further assessed and identified Packpred prediction accuracy for native-mutant pairs across different depths. Another area of improvement is the clique statistical potential that has many tunable parameters such as the number of amino acids in the clique, cut off distance and definitions of what constitutes a ‘contact’ between residues, etc. Packpred could improve by investigating these aspects too and this would form an independent study in itself. Similarly, further tweaks to the FADHM matrix, as briefly discussed above, could also possibly improve overall prediction accuracy. Shannon entropy accounts for the degree of variation at a given site/position, and does not change depending on the type of mutation. In our method, we use Shannon entropy in conjunction with the clique potential and FADHM to get a wholesome picture of sequence and structure conservation. However, it is likely that a more nuanced version of the entropy measure and/or other scores for conservation may help get more accurate predictions. In its current implementation, Packpred categorizes mutations as being neutral or destabilizing. When we tried to correlate the score with a discretized value of the function, the correlations were around -0.5. Perhaps, with some of the improvements discussed above this correlation would also improve.



One important observation from our findings is that of the 4099 mutations, 4036 were correctly called by at least one of the methods. There exists great complementarity between the methods tested here. We were tempted to then use two simple meta prediction methods. We designated the predictions involving a particular wild type amino acid or a wild type-mutant amino acid pair to the method that best predicted this type. Such a simple minded approach gave us MCCs of 0.40 and 0.51 for the amino acid and the amino acid pair type predictions respectively, where the best predicting method, Packpred, had an MCC of 0.36 (Missense3D data set). It is conceivable that a different method of combining the results from these different methods could vastly increase the accuracy of predicting the functional fate of single amino acid changes.

We assessed the robustness of Packpred by training it on the T4 set and a combination of the T4 and CcdB saturation mutagenesis data sets. Somewhat surprisingly, each of the training sets gave us different optimal values of feature weights. These different weights did not however affect the overall performance of the method on the Missense3D testing set. In earlier results too, we had observed that different weight combinations gave rise to similar performances on the training set. We believe that one of the primary reasons for the different optimal weights is the fact the three features in Packpred do not all affect predictions at the same level of granularity. The statistical potential and the substitution matrices (FADHM) give a score for particular mutations. Whereas, the Shannon entropy score gives a single value for a position, regardless of the type of mutation. Given the myriad of different environments and levels of conservation in different positions of the protein, the contribution due to each of these features is not uniformly the same across a protein. The positive aspect of these predictions is that, despite the lack in consensus of optimal values of the different features, the overall prediction accuracy does not appear to suffer. This is probably indicative of the fact that the features of the algorithm are important and perhaps a different way of combining these features may yield consistently better results.

In this chapter, we have presented our work on predicting the effects of mutation using a linear equation based approach. Specifically, we have used only 3 features in the linear equation. We took this simplistic approach to be better able to reason about the effect of mutations. However, it is possible that methods that use higher order equations may perform better. Including more features and then using machine learning based methods could be one of the things tried in the future. In addition to increasing the features, the training dataset can also be increased. However, currently we have only limited data for training the machine learning / deep learning models.

## Chapter 5

# Designing putative inhibitory small molecules against the Nipah virus proteins

- Selection of small drug-like molecules for docking
- Selection of target proteins from the Nipah virus
- Virtual screening using two distinct software
- Jury system for higher confidence in the putative binding small molecules
- MD simulations and binding free energy calculations to support the findings

The docking studies were done by Tejashree R. Kanitkar. The analysis, MD simulations and binding free energy calculations for the small molecule inhibitors were done in collaboration with Neeladri Sen and are a part of his thesis also.

Published - Sen N\*, **Kanitkar TR\***, Roy AA\*, Soni N, Amritkar K, Supekar S, Nair S, Singh G, Madhusudhan MS. Predicting and designing therapeutics against the Nipah virus. PLoS Negl Trop Dis. 2019 Dec 12;13(12):e0007419. doi: 10.1371/journal.pntd.0007419. PMID: 31830030; PMCID: PMC6907750.

[\* equal contributions]

## 5.1 Introduction

Nipah virus(NiV) is a RNA virus belonging to the genus Henipavirus of the Paramyxoviridae family[121]. The RNA is non-segmented and encodes nine proteins. Six of the nine proteins are structural proteins, namely, glycoprotein (G), fusion protein (F), phosphoprotein (P), nucleoprotein (N), matrix protein (M), and the RNA polymerase (L). The remaining three, C, V, and W are functional proteins. The virus attaches itself to the ephrine receptors of the host cell through its G protein and the fusion is mediated by the F protein. The F protein changes its conformation from the pre-fusion state to the post fusion state that results into the fusion of the virus with the host cells[122]. The N protein and P protein form the NP complex that binds with the viral RNA to form nucleocapsid that is used for transcription and replication of the virus. The M protein plays an important role in the assembly of newly generated proteins into viral particles. The M protein migrates towards the host cell membrane marking the location of budding and recruiting other necessary components to this site[123]. The three non structural proteins, V, W and C are all derived from the gene of P protein either by editing the RNA or by using an alternate open reading frame. All of these proteins are known to participate and inhibit the interferon signaling[124].

The initial outbreaks of the Nipah Virus(NiV) were reported in Malaysia and Singapore in 1999. The outbreak was observed in pigs and humans. Later, outbreaks were also reported in Bangladesh and India[125]. The most recent outbreak was seen in 2018, in an Indian state of Kerala[121], where the virus claimed the lives of 21 infected people. NiV has a high mortality rate of more than 70%, making it essential for designing therapeutics that would inhibit the viral proteins and the virus.

Small molecules that bind to the target proteins are computationally predicted using docking softwares that allows sampling of thousands of small molecules within a small amount of time, through a process called as virtual screening. Such virtual screens play a crucial role in the process of drug development. Broadly, docking softwares performs two steps to predict the putative binding molecules, first, sampling various conformations or poses of the small molecule and second, scoring the generated pose with respect to the target protein. The sampling step of these docking softwares use a variety of approaches like MD simulations, Monte Carlo methods, genetic algorithms, fragment based methods and many others[22]. The docking methods have shown tremendous success and potential for their application to other biomolecules[126].

In this study, we attempted to predict drug-like small molecules that could potentially bind and

inhibit the activity of the NiV proteins. We performed virtual screening of a library of drug-like small molecules on five of the NiV proteins for which X-ray crystal structures or good quality models were available. We used 2 different software for the virtual screens, giving us higher confidence in our results. Additionally, we performed MD simulations of the top scoring small molecules and calculated the binding free energy to support our findings.

## 5.2 Methods

### 5.2.1 Prediction of putative small molecules that can bind to NiV proteins:

Docking was used to identify putative small molecules that can potentially bind and inhibit the activities of the NiV proteins. In this exercise, NiV proteins (G, N, F, P and M proteins) that had crystal structures or models built from templates with high identity (>90%) and high coverage (>80%) were used as targets for ligand screening. The screening library consisted of a 70% non-redundant set of 22,685 ligands constructed from ~13 million clean drug like molecules of the ZINC database. The screening library consisted of 22685 ligands that were the 70% non-redundant set of ~13 million clean drug like molecules of the ZINC database[127, 128]. The 70% library was chosen as a practical measure to ensure wide coverage. Further, we envisage that during experimental trials all structurally similar small molecules to our predicted hits would be tested. The binding pockets for docking on the targets were predicted using the DEPTH server[42, 119]. The parameters of DEPTH included a minimum number of neighborhood waters set to 4 and the probability threshold for binding site of 0.8. Evolutionary information was also included by the server in binding site prediction. The druggability of the binding pocket was predicted using PockDrug[129] and CavityPlus[130], but no consensus prediction could be obtained (Table 1). Hence the druggability of the pocket was not taken into consideration during docking. Docking was performed using Autodock4[18], and DOCK6.8[19]. The target proteins were prepared for docking by Autodock4, by adding missing polar hydrogen atoms and Gasteiger charges. The ligand docking site, marked by affinity grids, was generated using the Autogrid module of Autodock. The center of the grid, number of grid points in X, Y, and Z directions and separation of grid points were chosen based on the predicted binding pockets using the ADT viewer from MGL tools[18]. The number of Genetic Algorithm runs was set to 20. The final energies reported by Autodock4 were used for evaluation and selection of the putative leads. The target proteins were prepared for docking by DOCK6.8 using Dock Prep tool [19] from

Chimera[131]. Missing hydrogen atoms were added to the target proteins using Chimera. Charges on atoms of the protein were determined using AMBER. Molecular surface of the target was generated using the DMS tool from Chimera. The sphgen program from DOCK6.8 was used to generate spheres from the molecular surface. The cluster of spheres were selected according to the binding sites predicted by DEPTH. The grid box and grid were created by showbox and grid programs respectively. Flexible ligand docking was performed using DOCK6.8. The final energies reported by DOCK6.8 were used for evaluation and selection of the putative leads.

### 5.2.2. Accessing the stability of small molecules against the NiV proteins:

The 13 small molecules were predicted with high confidence to bind different NiV proteins. Details of the procedures for modeling/predicting small molecule inhibitors are stated in the results section. MD simulations were carried out in triplicates for all four predicted protein-peptide inhibitor complexes. The simulations were carried out using GROMACS[26, 80] with the Amber99SB-ILDN force field[132]. Parameters for the small molecules were generated using Antechamber[133]. The Amber99SB-ILDN force field has been used for the MD simulations of protein-peptide and protein-ligand complexes extensively[134–136]. In an earlier study, we used the same force field to study various protein-ligand interactions and validated one such purported complex experimentally[137]. In the cases where the small molecule ligand dissociated from the binding site, we re-simulated the system using the CHARMM27 force field, another popularly used molecular mechanics package. We did the second simulation to ascertain that binding was indeed weak. Parameters for the small molecules in the CHARMM27 simulations were generated using SwissParam[138].

A water box whose sides were at a minimum distance of 1.2 nm from any protein atom was used for solvating each of the systems. Sodium or chloride counter ions were added to achieve charge neutrality. Electrostatic interactions were treated using the particle mesh Ewald sum method[82] and LINCS[139] was used to constrain hydrogen bond lengths. A time step of 2 fs was used for the integration. The whole system was minimized for 5000 steps or till the maximum force was less than 1000 kJ/mol/nm. The system was then heated to 300K in an NVT ensemble simulation for 100 ps using a Berendsen thermostat[83]. The pressure was stabilized in an NPT ensemble simulation for 100 ps using a Berendsen barostat. The systems were simulated (NPT) for a maximum of for 50 ns where pressure was regulated using the Parrinello-Rahman barostat[84].

Structures were stored after every 10ps. The temperature, potential energy and kinetic energy were monitored during the simulation to check for anomalies.

Free energy of binding of the putative small molecules provides an important quantitative description of its efficacy. In this study, the extensive MD simulations of protein-inhibitor complexes were post-processed to obtain binding free energy estimates using the molecular mechanics Poisson-Boltzmann surface area (MM/PBSA) approach[140, 141]. The MM/PBSA method employs an implicit solvation model to estimate the free energy of binding by evaluating ensemble averaged classical interaction energies (MM) and continuum solvation free energies (PBSA) of the protein-ligand complex conformations from the MD trajectories. The MM/PBSA calculations of the protein-small molecule inhibitors were calculated based on the last 40 ns trajectory with snapshots obtained after every 1000 ps, totalling to 40 snapshots. The MD snapshots were energy minimized for 2000 steps before evaluation of interaction and solvation free energies. The protein and solvent were modeled with dielectric constants of  $\epsilon = 2$  and  $\epsilon = 80$ , respectively. APBS suite[142] and GMXPBSA[143] were used for implicit solvent calculations. In this study, we attempted to calculate the entropic estimate of binding using the interaction entropy formalism[144]. However, converged entropic values with reasonable error estimates for the trajectories could not be obtained, which is often the case when evaluating entropic contributions from molecular simulations. We, therefore, neglected entropic contributions to the binding free energies, as estimated entropy change upon binding is often negligible and can be ignored for relative binding free energies calculations[141]. The enthalpies of binding obtained from MM/PBSA calculations are reported as binding energies for the complexes.

## 5.3 Results

### 5.3.1 Prediction of putative small molecules that can bind to NiV proteins:

The crystal structures of the G, N, P and F proteins were used in docking studies to find plausible small molecule inhibitors. A homology model of the M protein was also included in the docking exercise as it was based on a template with high (94%) sequence identity and coverage (88%). We were conservative with the docking approach and did not use our models of the structures of the W and V proteins in this exercise. Even though V and W proteins share a large portion of their

sequence with P, there was no crystal structure of P corresponding to the identical regions of V and W proteins (except residue no 1-38, which is too small a stretch for binding site prediction). The V and W protein models cover ~60% of the whole protein length (297 and 266 residues of a total length of 456 and 450 for V and W respectively) in discontinuous fragments, sometimes with target-template sequence identities of ~30%.) (<http://cospi.iiserpune.ac.in/Nipah/>)

First, we predicted the plausible binding pockets on each of the proteins using the DEPTH server that we had earlier benchmarked for binding site prediction accuracy. A total of 12 binding pockets were predicted in G (2), N (4), P (2), F (1) and M (3) proteins (Table 1). Two of the predicted binding pockets, one on the M protein and another on the G protein, are on the dimer interface and host protein (ephrin receptor) binding interface respectively. As mentioned in Methods section previously (Section 5.2), these sites are important drug targets. All 12 binding sites were used to screen 22685 drug like molecules from the 70% nonredundant ZINC database of clean drug like molecules using two different docking tools, DOCK6.8 and Autodock4. The docking tools provide a docking energy score that was used to select possible high affinity binders. In the absence of an objective measure or threshold to determine strong binders, we chose the top 150 best scoring ligands for each of the pockets from both the docking tools. We then compared the two lists for common molecules. 146 molecules were identified by both Dock6.8 and Autodock4 for G (9), N (56), P (45), F (10) and M (46) proteins (Table 2). The grid scores for the predicted complexes range between -71 to -32 units for DOCK6.8. The corresponding Autodock4 binding free energies range between -14 kcal/mol to -6 kcal/mol (Table 2).

Table 1: List of pocket lining residues for each pocket of NiV Proteins. The residue name is followed by the residue number. The chain id has been depicted after the dot.

NiV protein	Pocket	Pocket lining residue numbers	PockDrug	CavityPlus
Glycoprotein	PG1	F458.A, W504.A, Q559.A, D219.A, Y280.A, L305.A, Q490.A	0.43	Druggable
	PG2	P500.A, G489.A, R435.A, W479.A, S432.A, E430.A, R344.A, K376.A, F375.A, N378.A, S398.A, P383.A	0.47	Less druggable

Nucleoprotein	PN1	S67.A, A65.A, V58.A, I131.A, L128.A, E124.A, R36.A, F38.A, K34.A	0.99	Less druggable
	PN2	K69.A, N219.A, Q223.A, S224.A, L225.A, K229.A, F230.A, I35.A	0.99	Undruggable
	PN4	R218.A, N219.A, S222.A, R228.A, Q319.A, E316.A, I176.A, K178.A	0.99	Undruggable
	PN5	R307.A, Y310.A, V232.A, L314.A, E315.A, S226.A, D94.A, E233.A, L225.A	0.99	Undruggable
Phosphoprotein	PP1	T562.B, K559.B, V556.B, N561.C, T562.C, T566.C, E568.C, I567.C	0.43	Less druggable
	PP2	L517.C, E514.C, V516.C, N522.C, D482.B	Pocket not identified	Undruggable
Fusion protein	PF2	V39.B, Y30.B, H29.B, Y432.B, L433.B, N380.B, K40.B	0.88	Undruggable
Matrix protein	PM1	E195.A, H238.A, P332.A, Q328.A, L207.A, M236.A, D304.A, M188.A	0.33	Druggable
	PM2	F151.A, K143.A, W141.A, Y62.A, L181.A, Y187.A, M188.A, L274.A, D304.A	0.96	Druggable
	PM3	L312.A, W314.A, L309.A, F235.A, D213.A, M236.A, F266.A	0.96	Less druggable

To corroborate our predictions, we measured the RMSD between the same ligand (in the common list) as docked by the two different tools (top 5 poses predicted by Autodock4 were compared to the top pose predicted by DOCK6.8), after superimposing the proteins. This measure is referred to as RMSD<sub>lig</sub>. 15 unique drug-like molecules had an RMSD<sub>lig</sub> of less than 0.15 nm between their docked poses.



Table 2: List of the ranks and energy values of the small drug like molecules that were predicted in the top 150 scoring models by both DOCK6.8 and Autodock4. RMSD\_1 –RMSD\_5 are the RMSDs of the 5 best Autodock4 poses with the best scoring Dock6.8 pose. The least RMSD is depicted in bold. Pocket number indicates pockets from Autodock4. Some of the Autodock4 pockets have been subdivided by DOCK6.8, which indicates the subsections in each pocket

Protein name	PDB ID	Pocket Number	Number of selected molecules	ZINC ID	Rank in DOCK	Rank in Autodock4	Energy in DOCK6.8	Energy in Autodock	RMSD_1 (nm)	RMSD_2(nm)	RMSD_3(nm)	RMSD_4(nm)	RMSD_5(nm)
Glycoprotein	3D11	PG1	4	ZINC63411510	10	60	-58.1569	-8.73	0.8	<b>0.797</b>	0.809	0.807	0.797
		PG1		ZINC93305816	145	129	-48.6263	-8.55	<b>0.799</b>	0.8	0.803	0.876	0.861
		PG1		ZINC63857604	108	56	-49.4608	-8.76	<b>0.695</b>	0.698	0.697	<b>0.695</b>	0.697
		PG1		ZINC72264974	124	120	-49.0663	-8.57	1.033	<b>0.986</b>	<b>0.986</b>	1	1.005
		PG2	5	ZINC04580552	132	7	-42.6248	-8.97	2.26	<b>2.259</b>	2.293	2.312	2.293
		PG2		ZINC23214639	124	40	-42.7004	-8.43	2.395	2.373	2.393	2.349	<b>2.348</b>
		PG2		ZINC65407076	128	23	-42.6745	-8.64	<b>2.209</b>	2.211	2.232	2.225	2.246
		PG2		ZINC93655460	75	128	-43.7118	-8.09	2.271	2.267	<b>2.263</b>	2.271	2.265
		PG2		ZINC94217163	60	150	-44.5727	-8.04	2.295	2.293	<b>2.287</b>	2.3	2.306
Nucleoprotein	4CO6	PN1	1	ZINC34083937	138	74	-35.9687	-8.27	1.262	1.26	1.252	1.254	<b>1.239</b>
		PN1	10	ZINC42750806	48	50	-36.7541	-8.41	0.432	0.397	0.648	0.263	<b>0.251</b>
		PN1		ZINC02511792	62	62	-36.4557	-8.33	0.333	0.336	0.332	<b>0.32</b>	0.334
		PN1		ZINC34083937	52	73	-36.6097	-8.27	0.64	0.637	0.64	<b>0.601</b>	0.623
		PN1		ZINC05382414	68	147	-36.3915	-8.09	0.231	0.247	0.205	<b>0.179</b>	0.193
		PN1		ZINC16545537	107	5	-35.6182	-9.08	0.147	<b>0.145</b>	0.147	0.16	0.158
		PN1		ZINC16954338	141	15	-35.1782	-8.81	0.625	0.482	0.602	0.125	<b>0.124</b>
		PN1		ZINC63959595	25	139	-37.5303	-8.1	0.149	<b>0.142</b>	0.222	0.17	0.199
		PN1		ZINC92484162	140	86	-35.1959	-8.23	0.605	0.608	0.602	<b>0.6</b>	0.606
		PN1		ZINC92722391	142	40	-35.1454	-8.46	0.647	<b>0.621</b>	0.654	0.642	0.667

		PN1		ZINC92722539	125	20	-35.3924	-8.7	0.575	0.574	0.584	<b>0.268</b>	0.271
		PN2	15	ZINC94258465	24	96	-36.1295	-7.08	0.702	<b>0.693</b>	0.707	0.706	0.705
		PN2		ZINC94258558	33	79	-35.8008	-7.17	0.081	<b>0.074</b>	0.074	0.096	0.076
		PN2		ZINC86657759	64	54	-34.8881	-7.31	<b>0.505</b>	0.509	0.518	0.506	0.508
		PN2		ZINC73641145	6	28	-37.4871	-7.6	0.145	0.148	0.261	0.149	<b>0.142</b>
		PN2		ZINC95022396	98	42	-34.3319	-7.43	0.672	0.674	0.688	0.68	<b>0.661</b>
		PN2		ZINC77262630	4	50	-38.614	-7.35	0.275	<b>0.187</b>	0.232	0.221	0.255
		PN2		ZINC72264974	91	47	-34.3793	-7.36	1.115	1.094	<b>1.085</b>	1.103	1.171
		PN2		ZINC04580552	10	19	-36.8562	-7.79	0.375	0.376	0.286	<b>0.272</b>	0.442
		PN2		ZINC72107957	106	8	-34.1858	-8.16	0.508	<b>0.45</b>	0.455	0.761	0.722
		PN2		ZINC85191592	11	115	-36.8475	-7.02	0.685	0.691	0.71	0.747	<b>0.458</b>
		PN2		ZINC91932783	133	61	-34.0056	-7.26	0.173	0.173	<b>0.072</b>	0.167	0.503
		PN2		ZINC92349362	22	125	-36.1803	-6.99	0.323	0.353	0.362	<b>0.308</b>	0.334
		PN2		ZINC94927184	2	148	-39.6966	-6.92	0.647	0.657	0.656	<b>0.553</b>	0.659
		PN2		ZINC94937158	128	60	-34.0391	-7.28	<b>0.675</b>	0.699	0.733	0.7	0.72
		PN2		ZINC95355539	129	94	-34.0157	-7.1	0.778	0.77	0.785	0.765	<b>0.678</b>
		PN2	8	ZINC16755504	113	137	-36.6964	-6.96	1.962	<b>1.949</b>	2.367	2.359	2.368
		PN2		ZINC72129411	122	44	-36.5315	-7.41	<b>1.504</b>	1.516	1.531	1.525	1.51
		PN2		ZINC72133204	52	141	-37.8067	-6.94	1.843	<b>1.833</b>	1.843	1.836	1.844
		PN2		ZINC72388943	144	35	-36.2274	-7.52	1.889	1.89	1.895	<b>1.888</b>	1.896
		PN2		ZINC91932783	136	62	-36.285	-7.26	2.603	<b>2.602</b>	2.631	2.66	2.95
		PN2		ZINC94217163	147	87	-36.2132	-7.15	1.59	1.587	1.589	1.57	<b>1.301</b>
		PN2		ZINC94927184	2	150	-43.1395	-6.92	1.91	1.947	<b>1.869</b>	1.87	2.021
		PN2		ZINC95022396	132	42	-36.3742	-7.43	<b>2.006</b>	2.014	2.014	2.012	2.016

		PN2	11	ZINC14060343	38	128	-37.0963	-6.99	1.446	<b>1.427</b>	1.443	1.435	1.471
		PN2		ZINC35935889	135	2	-35.5389	-8.47	2.698	2.714	2.406	2.483	<b>1.551</b>
		PN2		ZINC72148214	115	55	-35.7789	-7.29	1.65	1.686	1.697	1.621	<b>1.498</b>
		PN2		ZINC72264974	114	46	-35.7853	-7.36	1.545	<b>1.501</b>	1.502	1.545	1.645
		PN2		ZINC94629031	103	112	-35.8935	-7.03	2.716	1.894	1.885	<b>1.875</b>	1.887
		PN2		ZINC94927184	29	150	-37.4455	-6.92	<b>1.542</b>	1.596	1.543	1.618	1.73
		PN2		ZINC73641145	61	28	-36.677	-7.6	1.368	1.373	<b>1.351</b>	1.356	1.369
		PN2		ZINC72129411	32	44	-37.2583	-7.41	<b>1.68</b>	1.693	1.696	1.686	1.682
		PN2		ZINC72107957	87	8	-36.2157	-8.16	<b>1.322</b>	1.401	1.418	1.538	1.529
		PN2		ZINC77262630	9	49	-38.7278	-7.35	<b>1.378</b>	1.431	1.442	1.432	1.418
		PN2		ZINC94937158	90	60	-36.0595	-7.28	1.436	1.409	<b>1.393</b>	1.404	1.401
		PN4	21	ZINC16932105	14	32	-40.6128	-9.07	0.917	0.917	0.92	0.533	<b>0.524</b>
		PN4		ZINC12362922	10	25	-41.366	-9.15	0.785	0.769	<b>0.139</b>	0.147	0.142
		PN4		ZINC92722391	45	82	-38.7467	-8.61	0.408	<b>0.394</b>	0.411	0.497	0.407
		PN4		ZINC00149964	57	29	-38.0889	-9.1	0.6	0.595	0.584	0.605	<b>0.577</b>
		PN4		ZINC06361369	81	24	-37.5147	-9.18	0.837	0.724	0.781	<b>0.718</b>	0.842
		PN4		ZINC02819777	65	5	-37.9025	-9.73	0.685	0.701	<b>0.65</b>	0.679	0.664
		PN4		ZINC04829362	21	97	-40.3007	-8.53	<b>0.085</b>	0.121	0.119	0.119	0.115
		PN4		ZINC04085190	39	33	-38.9491	-9.01	<b>0.467</b>	0.492	0.48	0.476	0.481
		PN4		ZINC00814199	8	64	-41.434	-8.77	0.553	0.51	0.514	<b>0.508</b>	0.526
		PN4		ZINC92179996	24	2	-39.9639	-9.82	0.637	0.64	0.638	0.613	<b>0.599</b>
		PN4		ZINC05378687	86	142	-37.3851	-8.36	0.291	0.338	0.327	0.299	<b>0.243</b>
		PN4		ZINC05603964	104	14	-36.9289	-9.38	0.772	0.745	0.746	<b>0.732</b>	0.752
		PN4		ZINC08913821	27	121	-39.8119	-8.42	0.773	0.777	0.702	0.855	<b>0.354</b>

		PN4		ZINC26481080	106	9	-36.9002	-9.49	0.488	0.47	0.491	0.486	<b>0.487</b>
		PN4		ZINC59209390	5	122	-43.6131	-8.42	0.368	0.36	<b>0.334</b>	0.336	0.378
		PN4		ZINC67489659	108	108	-36.8403	-8.45	0.26	0.245	<b>0.235</b>	0.26	0.274
		PN4		ZINC72165678	87	148	-37.3739	-8.34	0.187	<b>0.184</b>	0.201	0.205	0.195
		PN4		ZINC87440345	139	102	-36.3448	-8.5	0.748	0.737	0.724	0.734	<b>0.375</b>
		PN4		ZINC92722404	22	116	-40.2784	-8.42	0.312	0.319	0.293	<b>0.285</b>	0.382
		PN4		ZINC92722539	114	131	-36.7189	-8.4	0.762	0.507	<b>0.491</b>	0.531	0.603
		PN4		ZINC94725877	115	50	-36.7033	-8.86	0.261	0.274	0.275	0.266	<b>0.258</b>
		PN5	7	ZINC49587767	57	81	-36.7268	-6.59	0.655	<b>0.592</b>	0.631	0.67	0.62
		PN5		ZINC04334885	21	35	-37.8931	-6.83	0.744	0.742	0.736	0.762	<b>0.732</b>
		PN5		ZINC72107957	87	6	-36.2157	-7.65	0.664	0.66	0.588	<b>0.506</b>	0.524
		PN5		ZINC73641145	61	52	-36.677	-6.73	1.32	1.444	1.372	0.804	<b>0.741</b>
		PN5		ZINC35935889	135	2	-35.5389	-8.22	0.618	0.597	0.579	0.592	<b>0.575</b>
		PN5		ZINC72148214	115	67	-35.7789	-6.66	0.67	0.682	0.676	<b>0.615</b>	0.733
		PN5		ZINC95388070	106	145	-35.8458	-6.4	<b>0.955</b>	0.974	1.244	1.245	1.2
Phospho protein	4N5B	PP1	16	ZINC85650631	79	34	-34.2603	-6.71	0.168	<b>0.164</b>	0.233	0.232	0.167
		PP1		ZINC94927184	19	85	-35.8097	-6.55	<b>0.176</b>	0.19	0.177	0.395	0.385
		PP1		ZINC95384460	83	49	-34.1442	-6.65	0.327	0.333	0.33	<b>0.321</b>	0.326
		PP1		ZINC72462705	1	90	-39.3896	-6.54	0.195	0.167	0.13	0.178	<b>0.121</b>
		PP1		ZINC86098248	93	65	-34.0772	-6.6	<b>0.105</b>	0.108	0.109	0.108	0.107
		PP1		ZINC67884980	36	23	-34.97	-6.79	0.204	<b>0.195</b>	0.197	0.228	0.227
		PP1		ZINC77285117	38	41	-34.8839	-6.69	<b>0.144</b>	0.174	0.172	0.169	0.148
		PP1		ZINC95022396	52	89	-34.6239	-6.54	0.261	0.264	<b>0.256</b>	0.261	0.266
		PP1		ZINC13016500	105	144	-33.9586	-6.4	0.22	0.217	0.225	<b>0.215</b>	0.222

		PP1		ZINC24759441	128	149	-33.6575	-6.38	0.134	<b>0.133</b>	0.134	<b>0.133</b>	<b>0.133</b>
		PP1		ZINC32565459	97	138	-34.0351	-6.41	0.348	0.35	0.339	0.339	<b>0.338</b>
		PP1		ZINC65370580	70	147	-34.3435	-6.39	0.692	0.694	<b>0.687</b>	0.692	0.689
		PP1		ZINC65425676	65	139	-34.3686	-6.41	0.389	0.393	0.403	<b>0.379</b>	0.365
		PP1		ZINC77379208	125	16	-33.666	-6.87	0.58	0.582	0.588	0.586	<b>0.575</b>
		PP1		ZINC89195159	44	125	-34.7283	-6.44	0.33	0.343	0.328	0.343	<b>0.325</b>
		PP1		ZINC89201433	104	127	-33.9733	-6.44	0.21	0.207	0.201	<b>0.189</b>	0.209
		PP1	11	ZINC20534353	37	75	-34.0514	-6.57	0.31	0.305	0.301	0.303	<b>0.282</b>
		PP1		ZINC77379208	66	16	-33.3778	-6.87	<b>0.191</b>	0.215	0.214	0.2	0.197
		PP1		ZINC94927184	20	83	-35.1329	-6.55	0.589	0.564	0.588	0.594	<b>0.537</b>
		PP1		ZINC65405061	40	55	-33.9598	-6.62	<b>0.236</b>	0.239	0.239	0.24	0.242
		PP1		ZINC72462705	9	89	-36.1022	-6.54	<b>0.14</b>	0.227	0.2	0.229	0.198
		PP1		ZINC77285117	12	39	-35.8396	-6.69	0.177	0.136	0.152	<b>0.13</b>	0.18
		PP1		ZINC24759441	82	150	-33.1239	-6.38	0.144	0.144	0.144	0.144	<b>0.143</b>
		PP1		ZINC32565459	35	138	-34.2068	-6.41	<b>0.318</b>	0.33	0.32	0.32	0.319
		PP1		ZINC72133204	133	26	-32.4142	-6.75	<b>0.231</b>	0.233	0.234	<b>0.231</b>	<b>0.231</b>
		PP1		ZINC86098246	114	72	-32.6182	-6.58	0.234	<b>0.185</b>	0.208	0.226	0.21
		PP1		ZINC95448845	88	149	-33.0482	-6.38	0.53	0.572	<b>0.452</b>	0.509	0.547
		PP1	4	ZINC72462705	10	89	-62.9671	-6.54	7.624	<b>7.591</b>	7.638	7.599	7.632
		PP1		ZINC94927184	7	84	-64.3531	-6.55	7.708	7.709	7.716	7.587	<b>7.584</b>
		PP1		ZINC31394118	90	111	-56.0937	-6.48	7.653	7.661	7.661	<b>7.649</b>	<b>7.649</b>
		PP1		ZINC72438392	104	71	-55.1363	-6.59	7.514	7.522	7.517	7.522	<b>7.512</b>
		PP2	24	ZINC04722076	5	30	-68.4787	-9.91	<b>0.515</b>	<b>0.515</b>	<b>0.515</b>	<b>0.515</b>	<b>0.515</b>

		PP2		ZINC71260677	70	34	-56.161	-9.77	0.218	0.217	0.213	<b>0.208</b>	0.209
		PP2		ZINC94927184	7	49	-64.2946	-9.54	<b>2.032</b>	2.055	2.414	2.317	2.423
		PP2		ZINC67895025	62	92	-56.8376	-9.25	<b>0.66</b>	0.672	0.664	<b>0.66</b>	<b>0.66</b>
		PP2		ZINC19362297	74	67	-55.845	-9.41	0.301	<b>0.299</b>	<b>0.299</b>	0.314	0.299
		PP2		ZINC01584645	10	35	-62.4865	-9.76	<b>0.943</b>	1.031	1.027	1.014	1.073
		PP2		ZINC86094832	29	73	-59.68	-9.38	0.507	<b>0.346</b>	0.369	0.658	0.523
		PP2		ZINC86095599	34	97	-59.0738	-9.21	<b>0.118</b>	0.165	0.155	0.133	0.133
		PP2		ZINC92209154	35	29	-59.0716	-9.92	0.632	0.638	<b>0.62</b>	0.748	0.787
		PP2		ZINC95221243	47	33	-57.9412	-9.77	<b>0.949</b>	0.955	0.961	0.978	0.957
		PP2		ZINC91252717	2	1	-71.4697	-14.3	0.437	0.438	<b>0.427</b>	0.43	0.431
		PP2		ZINC35605802	38	15	-58.9016	-10.29	0.114	0.115	<b>0.111</b>	0.116	0.116
		PP2		ZINC72143751	91	27	-55.1678	-10.11	<b>2.07</b>	<b>2.07</b>	2.072	2.075	2.071
		PP2		ZINC72462705	14	24	-61.4933	-10.16	0.955	0.96	<b>0.866</b>	0.887	0.89
		PP2		ZINC19320365	132	16	-53.0723	-10.25	0.304	0.304	<b>0.303</b>	<b>0.303</b>	0.305
		PP2		ZINC19328716	59	105	-56.9384	-9.16	0.241	0.233	0.257	0.246	<b>0.212</b>
		PP2		ZINC36108799	122	109	-53.5489	-9.14	<b>0.249</b>	<b>0.249</b>	0.259	0.259	<b>0.249</b>
		PP2		ZINC39417829	48	113	-57.897	-9.13	0.422	<b>0.368</b>	0.496	0.501	0.498
		PP2		ZINC41206867	54	129	-57.1939	-9.03	<b>0.611</b>	0.69	0.65	0.679	0.685
		PP2		ZINC86094658	21	103	-60.3732	-9.19	0.458	0.444	0.421	0.543	<b>0.216</b>
		PP2		ZINC86662794	111	111	-53.9779	-9.14	0.239	<b>0.2</b>	0.202	0.234	0.236
		PP2		ZINC86680029	112	91	-53.9641	-9.25	0.503	<b>0.502</b>	<b>0.502</b>	<b>0.502</b>	0.513
		PP2		ZINC86730664	18	102	-60.9041	-9.19	0.954	0.956	0.946	<b>0.913</b>	<b>0.913</b>
		PP2		ZINC87254662	81	135	-55.7055	-8.99	0.594	0.593	<b>0.592</b>	0.599	0.606
Fusion protein	5EV M	PF1	3	ZINC19558876	118	141	-42.6413	-8.03	5.815	<b>5.791</b>	5.835	5.841	5.828
		PF1		ZINC44831966	140	23	-42.3345	-8.68	5.651	5.665	5.639	<b>5.62</b>	5.624

		PF1		ZINC63411510	15	34	-46.6526	-8.54	59.52	59.61	59	58.96	59.01
		PF2	7	ZINC94725877	91	43	-38.3918	-8.82	0.397	0.405	0.4	0.374	<b>0.311</b>
		PF2		ZINC72131030	74	4	-38.7232	-9.52	<b>0.45</b>	0.452	0.473	0.471	0.457
		PF2		ZINC93518195	68	92	-38.8302	-8.5	0.582	0.582	0.582	0.585	<b>0.578</b>
		PF2		ZINC65418720	7	94	-42.7932	-8.49	<b>0.541</b>	0.566	0.579	0.567	0.578
		PF2		ZINC34083754	65	35	-38.9075	-8.9	0.362	0.369	0.385	0.357	<b>0.343</b>
		PF2		ZINC00467624	52	127	-39.1657	-8.4	0.34	<b>0.313</b>	<b>0.313</b>	0.332	0.322
		PF2		ZINC04337208	141	11	-37.7391	-9.2	0.434	0.434	0.434	0.43	<b>0.425</b>
Matrix protein	Mono mer of modeled dimer	PM1	1	ZINC02511792	41	87	-39.7644	-7.45	0.543	0.503	0.599	0.504	<b>0.48</b>
		PM1	3	ZINC02819777	26	22	-36.5827	-7.85	1.689	1.668	<b>1.664</b>	1.674	1.683
		PM1		ZINC00344036	84	120	-32.7182	-7.34	0.709	<b>0.708</b>	0.71	0.71	0.71
		PM1		ZINC05603964	23	129	-37.1505	-7.3	1.941	1.953	1.955	1.924	<b>1.919</b>
		PM2	23	ZINC26481080	53	25	-45.3372	-8.55	0.415	0.394	0.414	0.389	<b>0.388</b>
		PM2		ZINC12362922	78	29	-44.4204	-8.51	<b>0.183</b>	0.186	0.185	<b>0.183</b>	<b>0.183</b>
		PM2		ZINC00814199	14	7	-49.4513	-8.97	0.624	0.621	0.622	<b>0.619</b>	0.624
		PM2		ZINC31165406	34	20	-46.4279	-8.61	0.4	0.392	<b>0.236</b>	0.381	0.47
		PM2		ZINC00149964	31	16	-47.1346	-8.67	0.624	0.621	0.622	<b>0.619</b>	0.624
		PM2		ZINC01725633	20	15	-48.5007	-8.68	0.409	0.42	0.204	0.399	<b>0.147</b>
		PM2		ZINC16932105	37	34	-46.151	-8.48	0.315	<b>0.314</b>	0.315	0.315	<b>0.314</b>
		PM2		ZINC71789643	73	30	-44.5767	-8.51	0.364	0.352	0.366	<b>0.341</b>	0.362



		PM2		ZINC93518353	97	100	-43.7746	-8.02	0.41	0.411	<b>0.397</b>	0.464	0.492
		PM2		ZINC91497887	87	67	-44.0917	-8.21	0.192	0.187	0.19	<b>0.185</b>	0.2
		PM2		ZINC02819777	90	28	-44.0287	-8.51	0.475	0.469	0.471	<b>0.414</b>	0.482
		PM2		ZINC19735365	86	93	-44.1054	-8.04	<b>0.529</b>	<b>0.529</b>	0.532	0.531	0.532
		PM2		ZINC04085190	40	113	-45.9788	-7.97	<b>0.352</b>	0.354	<b>0.352</b>	0.354	0.354
		PM2		ZINC04829362	79	107	-44.3996	-7.99	0.315	0.313	0.316	<b>0.312</b>	0.313
		PM2		ZINC05331903	144	22	-42.7511	-8.58	<b>0.463</b>	0.474	0.469	0.47	0.496
		PM2		ZINC05372521	48	139	-45.6329	-7.88	0.291	0.291	0.332	0.331	<b>0.286</b>
		PM2		ZINC05382414	29	103	-47.3167	-8.01	0.539	0.526	<b>0.438</b>	0.443	0.447
		PM2		ZINC20154773	113	36	-43.4471	-8.44	0.171	0.165	0.165	<b>0.164</b>	0.174
		PM2		ZINC22130393	139	30	-42.8524	-8.51	0.162	<b>0.156</b>	0.161	0.162	0.159
		PM2		ZINC45070221	52	130	-45.3486	-7.91	0.164	0.184	<b>0.142</b>	0.166	0.161
		PM2		ZINC63781317	115	23	-43.4275	-8.57	0.601	0.212	0.215	0.212	0.123
		PM2		ZINC72131030	147	45	-42.6896	-8.39	0.361	0.304	0.523	0.428	<b>0.295</b>
		PM2		ZINC92722404	74	143	-44.5125	-7.87	<b>0.507</b>	0.539	0.57	0.523	0.584
		PM2	26	ZINC26481080	20	24	-45.2565	-8.55	0.352	0.355	0.364	<b>0.293</b>	<b>0.293</b>
		PM2		ZINC93518353	36	98	-43.0793	-8.02	0.389	0.387	0.385	0.447	<b>0.35</b>
		PM2		ZINC00814199	27	7	-44.3983	-8.97	0.655	<b>0.645</b>	0.653	0.652	0.682
		PM2		ZINC00149964	12	16	-46.6062	-8.67	<b>0.452</b>	0.454	0.459	0.457	0.457
		PM2		ZINC02819777	92	29	-40.6191	-8.51	0.315	0.316	0.337	0.384	<b>0.302</b>
		PM2		ZINC31165406	22	20	-45.006	-8.61	0.424	0.42	<b>0.297</b>	0.398	0.446
		PM2		ZINC02511792	68	67	-41.305	-8.21	<b>0.165</b>	0.187	0.185	0.173	0.166
		PM2		ZINC91497887	89	66	-40.668	-8.21	<b>0.247</b>	<b>0.247</b>	0.254	0.253	0.252
		PM2		ZINC00129345	66	95	-41.371	-8.04	0.423	0.454	<b>0.41</b>	0.441	0.42
		PM2		ZINC01725633	58	15	-41.7507	-8.68	0.423	0.418	<b>0.415</b>	0.418	0.447

		PM2		ZINC04020772	23	46	-44.958	-8.37	0.457	0.461	0.452	<b>0.449</b>	0.461
		PM2		ZINC04085190	42	113	-42.8898	-7.97	0.494	0.496	0.495	0.495	<b>0.49</b>
		PM2		ZINC04829362	64	108	-41.4436	-7.99	0.218	0.226	0.216	<b>0.214</b>	0.227
		PM2		ZINC04962728	142	43	-39.4863	-8.4	0.55	0.552	0.55	0.551	<b>0.549</b>
		PM2		ZINC05382414	8	103	-48.0795	-8.01	0.537	0.518	0.416	0.41	<b>0.404</b>
		PM2		ZINC12362922	148	29	-39.3891	-8.51	0.387	0.388	0.387	<b>0.385</b>	0.386
		PM2		ZINC20154773	133	36	-39.7094	-8.44	1.956	1.95	<b>1.946</b>	<b>1.945</b>	1.964
		PM2		ZINC45070221	59	131	-41.6704	-7.91	0.389	<b>0.383</b>	0.4	0.39	0.393
		PM2		ZINC45796058	94	121	-40.5724	-7.95	0.734	0.734	0.739	0.738	<b>0.728</b>
		PM2		ZINC65407126	131	18	-39.7992	-8.62	1.666	1.666	<b>1.635</b>	1.642	1.644
		PM2		ZINC83236053	112	118	-40.1324	-7.96	0.629	0.581	0.318	0.322	<b>0.317</b>
		PM2		ZINC87017834	130	115	-39.816	-7.97	0.288	0.312	0.307	<b>0.285</b>	0.288
		PM2		ZINC92711149	119	77	-40.0153	-8.13	0.376	0.377	<b>0.375</b>	0.376	<b>0.375</b>
		PM2		ZINC92722404	33	144	-43.8065	-7.87	0.469	0.484	0.514	<b>0.462</b>	0.544
		PM2		ZINC94936845	32	106	-43.9738	-7.99	0.42	0.412	0.403	0.332	<b>0.306</b>
		PM2		ZINC95359457	103	109	-40.3436	-7.99	1.756	<b>1.744</b>	1.745	1.777	1.749
		PM3	10	ZINC49453727	86	70	-38.7917	-6.56	1.949	<b>1.94</b>	1.946	1.969	1.992
		PM3		ZINC83328368	25	33	-40.687	-6.82	1.775	1.781	1.775	<b>1.74</b>	1.743
		PM3		ZINC72131030	72	19	-38.9782	-6.9	1.892	<b>1.89</b>	1.914	1.893	1.905
		PM3		ZINC20390482	60	94	-39.1807	-6.48	1.821	1.821	1.821	<b>1.819</b>	1.82
		PM3		ZINC63781317	51	53	-39.3671	-6.66	1.839	1.829	1.829	<b>1.746</b>	1.796
		PM3		ZINC20163996	12	72	-42.2337	-6.56	1.896	1.896	1.894	<b>1.893</b>	<b>1.893</b>
		PM3		ZINC00189011	112	142	-38.2955	-6.38	<b>1.644</b>	1.648	1.649	1.65	1.653
		PM3		ZINC86864968	134	10	-38.0403	-7.23	1.744	1.705	<b>1.68</b>	1.723	1.78
		PM3		ZINC89949696	55	128	-39.2591	-6.4	1.941	1.944	1.944	1.494	<b>1.479</b>

		PM3		ZINC95008629	93	122	-38.6525	-6.41	1.973	1.977	<b>1.972</b>	1.978	1.975
		PM3	8	ZINC20163996	70	69	-33.027	-6.56	2.471	2.47	2.466	2.464	<b>2.463</b>
		PM3		ZINC05603964	23	27	-37.1505	-6.85	3.517	3.53	3.526	3.524	<b>3.515</b>
		PM3		ZINC72131030	48	19	-34.1461	-6.9	1.832	1.837	<b>1.793</b>	1.827	1.803
		PM3		ZINC91497887	9	43	-39.8472	-6.73	3.406	3.389	3.426	<b>3.384</b>	3.425
		PM3		ZINC73736970	88	24	-32.4629	-6.86	3.397	<b>3.339</b>	3.412	3.433	3.427
		PM3		ZINC02819777	26	15	-36.5827	-7.1	3.317	3.338	<b>3.303</b>	3.342	3.356
		PM3		ZINC01418749	113	133	-31.9688	-6.39	1.652	1.644	<b>1.641</b>	1.654	1.649
		PM3		ZINC33295102	21	121	-37.3533	-6.41	3.37	3.367	<b>3.365</b>	3.384	3.384

In addition to conformational similarity, we also assessed the similarities in ligand-protein interactions, primarily hydrogen bonding (Table 3). Further, the hydrogen bonding interactions were ~50 % conserved in 9 of these complexes (with RMSD\_lig < 0.15 nm). In a few instances, though the hydrogen bonding was not precisely the same, visual inspection of the complexes suggests that these bonds could be formed with small conformational changes.

Table 3: Number of hydrogen bonds that are formed between the selected pose for DOCK6.8 and Autodock4 with the protein. Number of common hydrogen bonds indicates the number of hydrogen bonds that are common between the predicted poses of the ligand from Autodock4 and DOCK6.8. \*\* The RMSD between DOCK and Autodock is 0.427 nm (greater than the cutoff). This entry is included as the rank for this ligand DOCK is 2 and Autodock is 1, indicating higher confidence in the prediction

Sr no	Protein Name	Pocket Number	ZINC ID	Number of Hydrogen bonds for DOCK6.8	Number of Hydrogen bonds for Autodock4	Number of common Hydrogen bonds
1	Nucleoprotein	PN21	ZINC94258558	4	3	3
2	Nucleoprotein	PN21	ZINC73641145	7	7	4
3	Nucleoprotein	PN4	ZINC12362922	5	7	2
4	Phosphoprotein	PP11	ZINC72462705	1	2	1
5	Phosphoprotein	PP11	ZINC86098248	1	1	0
6	Phosphoprotein	PP11	ZINC77285117	1	2	1
7	Phosphoprotein	PP12	ZINC72462705	1	2	0
8	Phosphoprotein	PP12	ZINC77285117	0	1	0
9	Phosphoprotein	PP2	ZINC86095599	1	1	0
10**	Phosphoprotein	PP2	ZINC91252717	3	2	0
11	Phosphoprotein	PP2	ZINC35605802	2	4	2
12	Nucleoprotein	P12	ZINC16545537	2	4	2
13	Nucleoprotein	P12	ZINC63959595	4	4	3

14	Nucleoprotein	PN21	ZINC91932783	3	5	3
15	Nucleoprotein	PN4	ZINC12362922	5	7	2
16	Nucleoprotein	PN4	ZINC04829362	5	6	5
17	Phosphoprotein	PP11	ZINC24759441	0	1	0
18	Phosphoprotein	PP11	ZINC77285117	0	2	0
19	Phosphoprotein	PP12	ZINC24759441	0	1	0
20	Phosphoprotein	PP21	ZINC86095599	1	1	0
22	Matrix protein	PM21	ZINC45070221	3	3	2
23	Matrix protein	PM21	ZINC01725633	4	5	2

10 drug-like molecules in N (4), P (5) and M (1) had an RMSD\_lig of less than 0.15 nm between their docked poses and were in the top 100 scoring models as predicted by both the docking tools. We did not get molecules that had a RMSD\_lig of less than 0.15nm for the G and F proteins. The molecule with the best RMSD\_lig (0.074 nm) from our screening, ZINC94258558 (Figure 1-A), binds the N protein (Table 2). Molecules ZINC73641145, ZINC12362922, and ZINC04829362 also have RMSD\_lig less than 0.15 and are predicted to bind to the N protein. ZINC73641145 has a DOCK6.8 rank of 6 and AutoDock rank of 28, indicating their better binding pose in comparison to the other sampled poses and small molecule ligands. Small molecules ZINC72462705, ZINC86098248, ZINC77285117, ZINC86095599 and ZINC35605802 are predicted binders to P that have RMSD\_lig of less than 0.15nm. Particularly, ZINC72462705 is the best scoring and top ranking molecule from the DOCK6.8 run while it scores rank 90 from the AutoDock run. Such examples need further investigation of how well the ligand really binds to a given protein pocket. For the M protein, ZINC01725633 was shortlisted (Table 2). There are however 3 molecules (Table 2) that are of interest despite their relatively large RMSD\_lig values. The molecule ZINC91252717 is predicted as the best binder to the P protein by Autodock4 (binding energy of -14 kcal/mol) and the second best binder by DOCK6.8 (grid score of -71) (Figure 1-B). These scores were among the best achieved during this docking exercise. We selected ZINC00814199 that was docked onto the M protein and was similar to ZINC01725633, which in turn formed 14 and 8 hydrogen bonds with Autodock4 and Dock6.8 respectively. ZINC00814199 was within the top 14 ranked compounds by both methods. Lastly, the hydrophobic molecule ZINC63411510 is predicted to bind the G protein on its ephrin-B2 binding

interface. Though both docking methods identified this site, the docking poses were different (RMSD<sub>lig</sub> of 0.8 nm). We hypothesize that the hydrophobic nature of the binding pocket and its size could contribute to the difference in docked poses. Note that in our list there are 3 ligands (ZINC12362922, ZINC00814199 and ZINC73641145) that (Table 4) bind different pockets on the same protein or pockets on different proteins. The ligand binding pockets (PN4 and PM2) that bind ZINC12362922 and ZINC00814199 have a similar amino acid composition containing Lys/Arg residues, Tyr residue and Leu/Val residues. The two ligands have terminal oxygens that interact with positively charged residues of the binding pocket. Another ligand ZINC73641145 binds to two different pockets on N protein (PN5 and PN4), these pockets are spatially close to one another and the ligand occupies the region between the two pockets in a similar orientation. Interestingly, a known drug (ZINC04829362), an antiasthmatic and antipsoriatic among other uses, binds to a pocket of the N protein with RMSD<sub>lig</sub> of 0.085 nm. Another drug (ZINC12362922) used in the treatment of depression and Parkinson's disease also binds the N protein with RMSD<sub>lig</sub> < 0.15 nm.

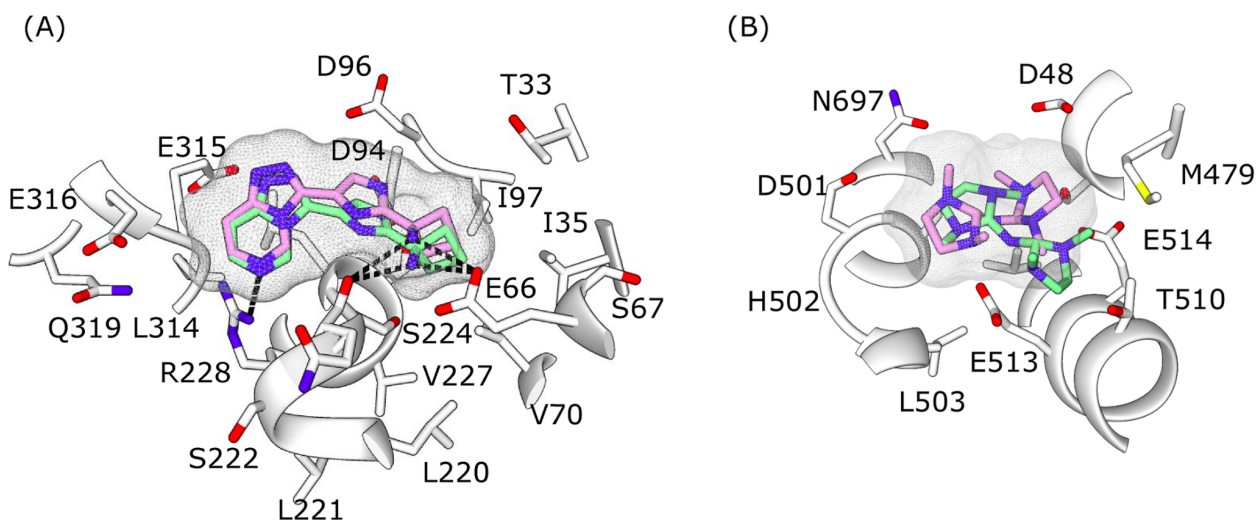


Figure 1: The docked poses of ZINC94258558 bound to N protein (A) and ZINC91252717 bound to P protein (B) as predicted by Autodock4 (green sticks with surface mesh) and Dock6.8 (lilac sticks with surface mesh). The protein is represented in white ribbons with the residues interacting with ligand shown in stick representation. Hydrogen bonds (only displayed in A) are shown as dashed lines.

Table 4: Same drug like molecule predicted to bind different pockets of the same or different protein. The binding pocket has been mentioned in parenthesis.

Sr. No.	ZINC ID	Pocket Number
1	ZINC72148214	PN23,PN5
2	ZINC05603964	PN4,PM12,PM32
3	ZINC93518353	PM21,PM22
4	ZINC04829362	PN4,PM21,PM22
5	ZINC77285117	PP11,PP11,PP12
6	ZINC92722404	PN4,PM21,PM22
7	ZINC16932105	PN4,PM21
8	ZINC34083937	PN11,PN12
9	ZINC12362922	PN4,PM21,PM22
10	ZINC00814199	PN4,PM21,PM22
11	ZINC72462705	PP11,PP12,PP13,PP2
12	ZINC95022396	PN21,PN22,PP11
13	ZINC94927184	PN21,PN22,PN23,PP11,PP12,PP13,PP2
14	ZINC72131030	PF2,PM21,PM31,PM32
15	ZINC20163996	PM31,PM32
16	ZINC32565459	PP11,PP12
17	ZINC02511792	PN12,PM11,PM22
18	ZINC94725877	PN4,PF2
19	ZINC63781317	PM21,PM31
20	ZINC94217163	PG2,PN22
21	ZINC24759441	PP11,PP12
22	ZINC91932783	PN21,PN22
23	ZINC20154773	PM21,PM22
24	ZINC91497887	PM21,PM22,PM32
25	ZINC77262630	PN21,PN22
26	ZINC45070221	PM21,PM22
27	ZINC31165406	PM21,PM22
28	ZINC26481080	PN4,PM21,PM22
29	ZINC04580552	PG2,PN21
30	ZINC72264974	PG1,PN21,PN23
31	ZINC63411510	PG1,PF1

32	ZINC05382414	PN12,PM21,PM22
33	ZINC02819777	PN4,PM12,PM21,PM22,PM32
34	ZINC72133204	PN22,PP12
35	ZINC72129411	PN22,PN22
36	ZINC35935889	PN23,PN5
37	ZINC72107957	PN21,PN22,PN5
38	ZINC77379208	PP11,PP12
39	ZINC92722391	PN12,PN4
40	ZINC94937158	PN21,PN22
41	ZINC92722539	PN12,PN4
42	ZINC73641145	PN21,PN22,PN5
43	ZINC04085190	PN4,PM21,PM22
44	ZINC00149964	PN4,PM21,PM22
45	ZINC01725633	PM21,PM22

### 5.3.2 Computational prediction of the stability of the protein-inhibitor complexes:

To assess the stability of the 13 protein-small molecule ligand complexes, we carried out three independent MD simulations of 50 ns each, using the AMBER99SB-ILDN force field. Simulations were carried out for 10 of the 13 ligands that had RMSD<sub>lig</sub> less than 0.15nm (Table 2) starting with the DOCK6.8 predicted pose. For each of the trajectories, the distance of the centre of the small molecule ligand to the centre of the binding pocket (based on the starting structure after NPT equilibration) was monitored (Figure 2). The triplicate MD simulations were terminated if this distance in 2 of the 3 trajectories exceeded 1 nm from its starting value. This happened in 5 cases, 2 inhibitors from N and P each and 1 from the M protein. These complexes were then re-simulated using the CHARMM27 force field. To summarize, we found that 2 inhibitors of N and 4 inhibitors of P showed stable binding in either AMBER99SB-ILDN or CHARMM27 MDs. For the 3 ligands with RMSD<sub>lig</sub> > 0.15 nm simulations were carried out starting with both the DOCK6.8 and Autodock4 predicted poses. Of these 3 ligands, the one that bound to G, showed stable association with it as quantified by the distance less than 0.6nm in all the triplicates of the CHARMM27 run. Similarly, the inhibitors of P and M (that had RMSD<sub>lig</sub> > 0.15nm) also showed stable associations.

To further our confidence in the predicted inhibitory molecules, we computed binding energies



for the protein-ligand complexes using MM/PBSA. 9 of the binding energies were computed to be negative in at least one of the replicates (3 for N protein, 4 for P protein, 1 for G protein and 1 for M protein). In one case (P protein-ZINC7262705 ligand), the binding energy with the CHARMM force field (after the AMBER simulation was terminated) was computed to have positive binding free energy (Table 5 and 6). In 3 cases (1 for N, P and M protein each) the ligand did not remain bound to the protein in either CHARMM or/and AMBER simulations.

The two known drugs, ZINC04829362 and ZINC12362922 remained bound to the N protein in all 3 replicates with negative binding energies in at least 2 of the trajectories. For the important druggable site on the G protein, the ligand remained bound in all 3 replicates when starting with the Autodock4 bound pose with negative binding energies

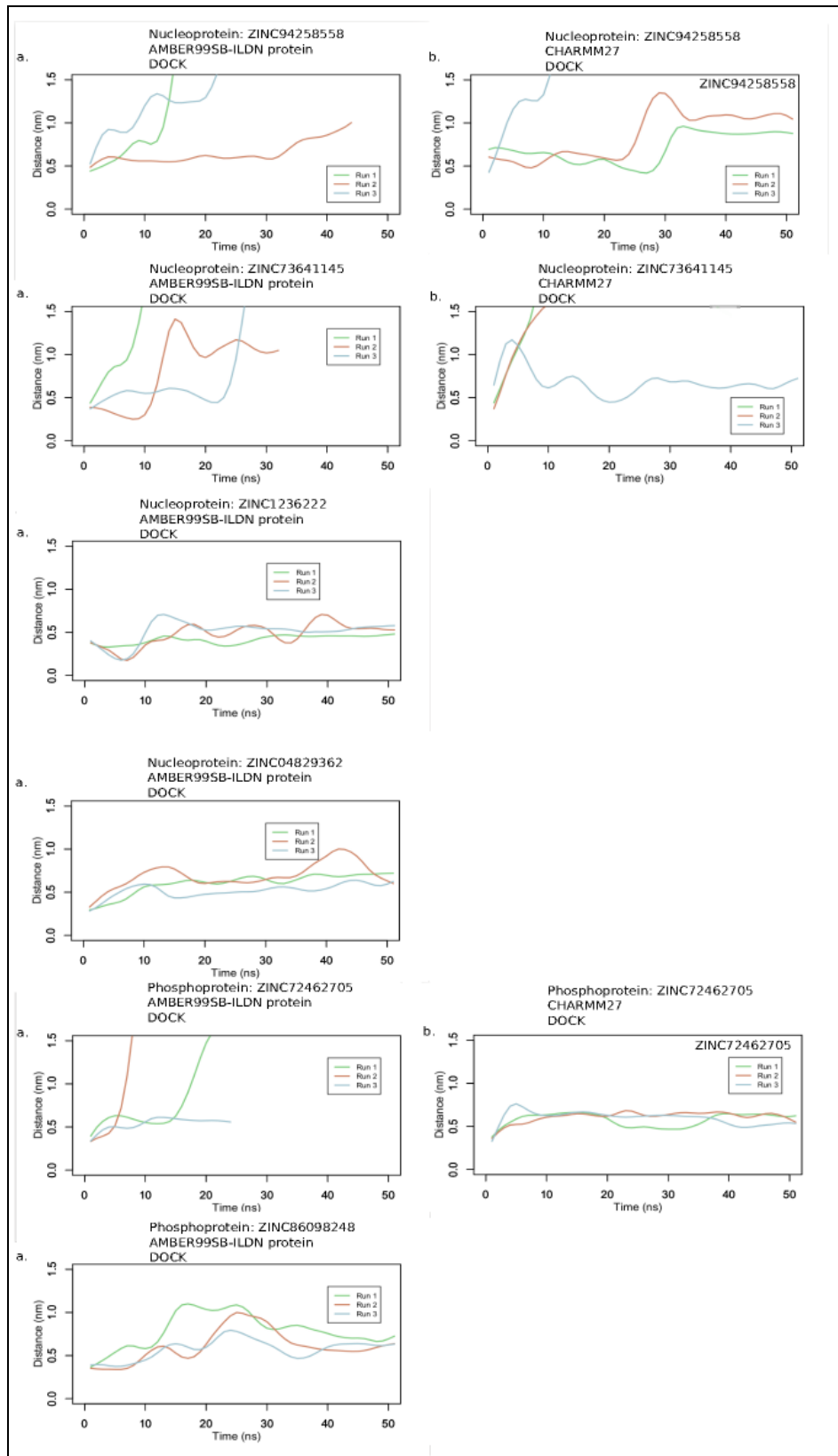
Table 5: Binding free energy as predicted using MM/PBSA calculations from molecular dynamics simulations carried out using AMBER and CHARMM force fields for 10 ligands predicted against N, P and M proteins. The binding free energies were not calculated (depicted by -) when the ligand left the binding site in at least 2 out of 3 replicates. CHARMM was only used to run molecular dynamics simulations when the ligand left the binding pocket in AMBER simulations.

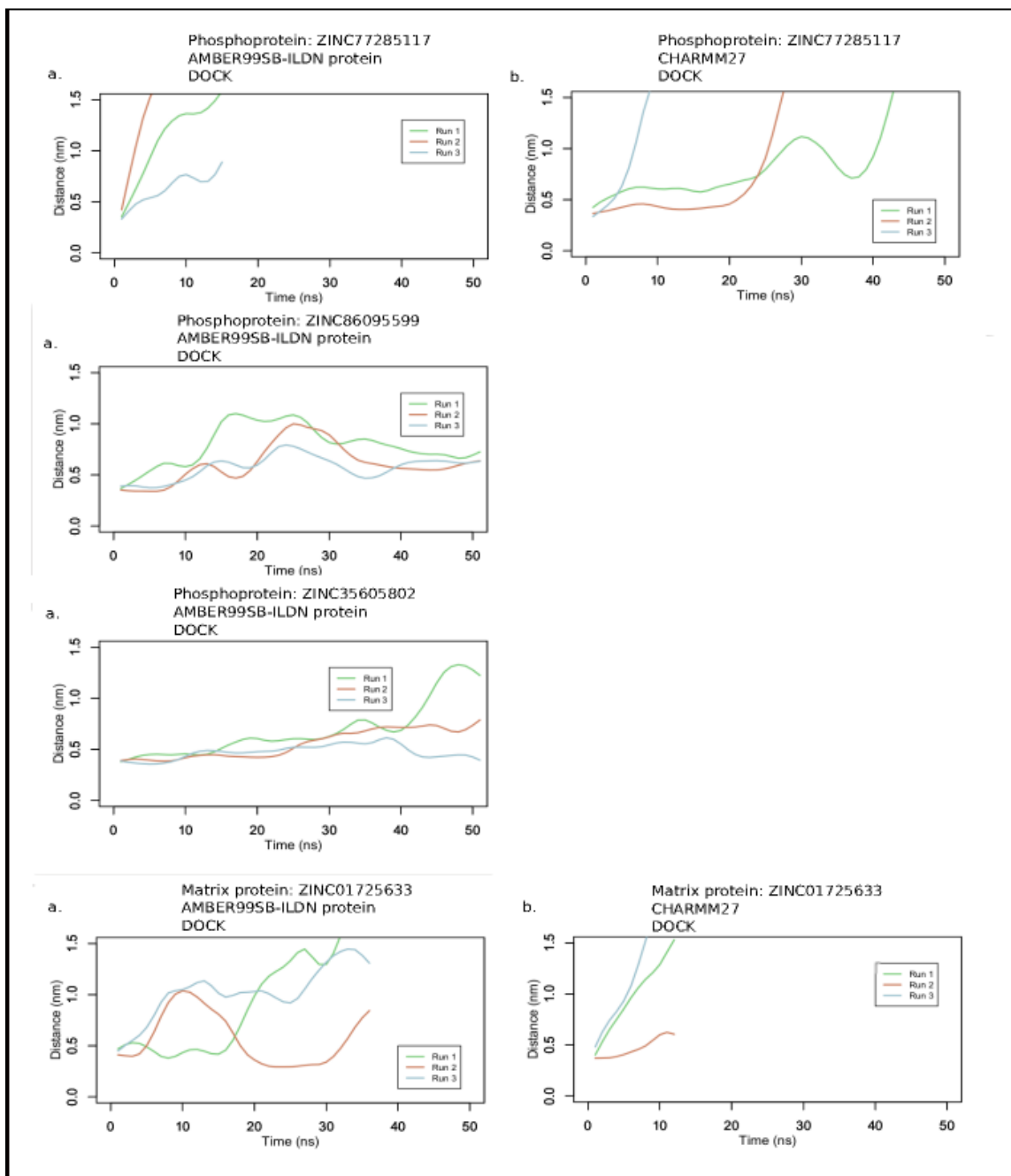
ZINC ID	Protein	Replicate	Binding free energy as predicted during	
			AMBER simulation (kJ/mol)	CHARMM simulation(kJ/mol)
ZINC94258558	N	1	-	-114+/-10
		2	-	-86+/-4
		3	-	-
ZINC73641145	N	1	-	-
		2	-	-
		3	-	-
ZINC12362922	N	1	-96+/-8	

		2	-100+/-10	
		3	-69+/-6	
ZINC04829362	N	1	-37+/-7	
		2	86+/-7	
		3	-101+/-8	
ZINC72462705	P	1	-	106+/-4
		2	-	86+/-4
		3	-	98+/-5
ZINC86098248	P	1	39+/-5	
		2	-65+/-7	
		3	37+/-3	
ZINC77285117	P	1	-	-
		2	-	-
		3	-	-
ZINC86095599	P	1	-196+/-7	
		2	-149+/-10	
		3	-153+/-14	
ZINC35605802	P	1	-14+/-0.5	
		2	1+/-1	
		3	-98+/-8	
ZINC01725633	M	1	-	-
		2	-	-
		3	-	-

Table 6: Binding free energy as predicted using MM/PBSA calculations from molecular dynamics simulations carried out using AMBER force fields for 3 ligands predicted against G, M and P proteins for both the predicted DOCK6.8 and Autodock4 poses. The binding free energies were not calculated (depicted by -) when the ligand left the binding site in at least 2 out of 3 replicates.

ZINC ID	Protein	Replicate	Binding free energy as predicted from (kJ/mol)	
			DOCK pose (kJ/mol)	Autodock pose (kJ/mol)
ZINC00814199	M	1	-153+/-6	-119+/-8
		2	-	-184+/-3
		3	-203+/-6	-
ZINC63411510	G	1	-	-44+/-4
		2	-	-79+/-4
		3	-	-59+/-4
ZINC91252717	P	1	-158+/-9	-187+/-8
		2	-256+/-10	-196+/-8
		3	-251+/-7	-





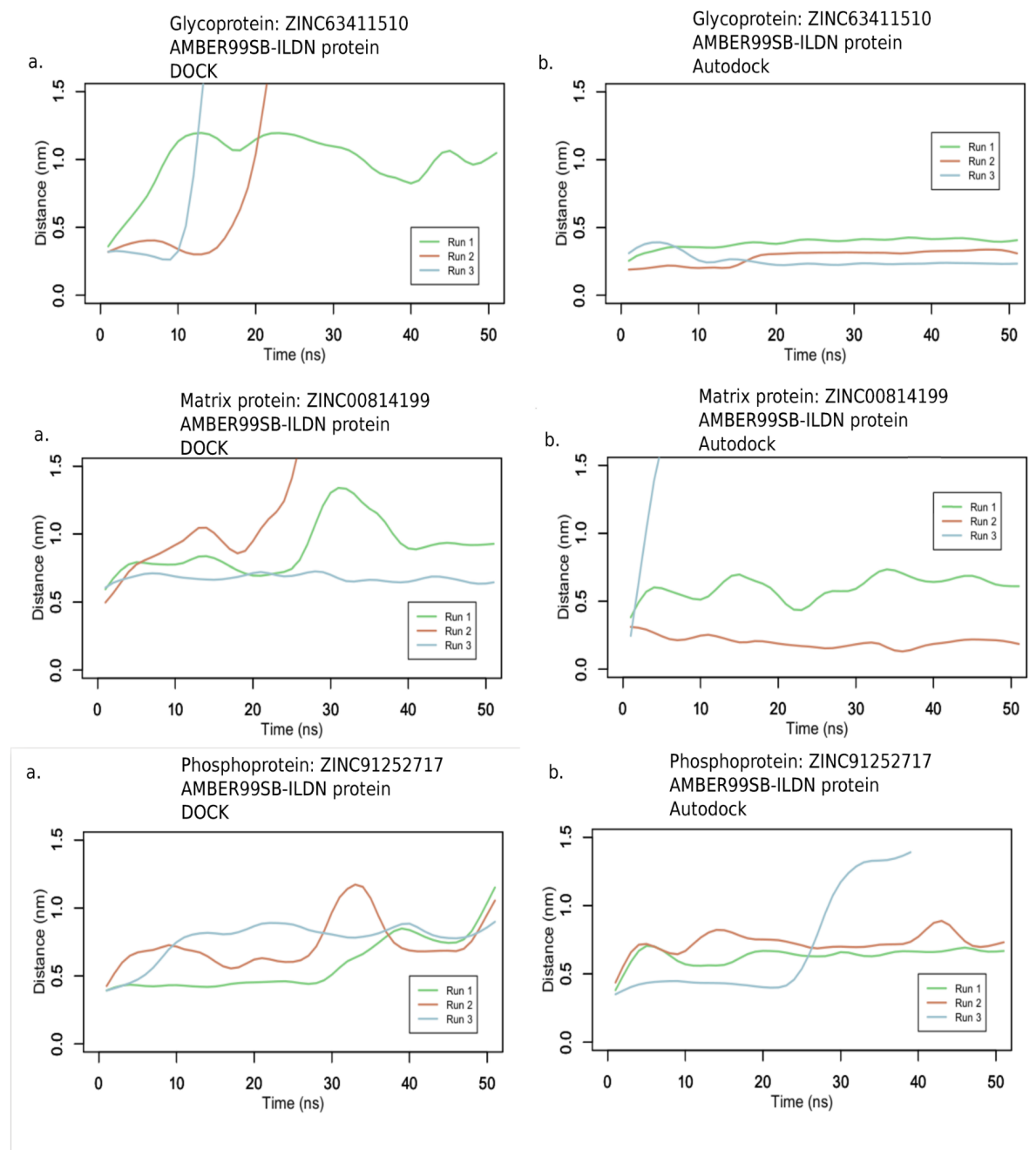


Figure 2: Distance of the center of the ligand from the center of the binding site (calculated based on the residues within  $5\text{\AA}$  of the first snapshot after NPT equilibration) during the simulation. The identity of the ligand, force field and docking strategy used and the target protein has been indicated above each plot.

## 5.4 Discussions

Nipah is a deadly virus with a mortality rate higher than 70%. Despite this, there are no approved drugs against NiV. In this study, we attempted to predict putative small drug-like molecules that could bind to various NiV proteins and inhibit their activity. Towards this, we virtually screened ~22000 molecules from the ZINC library against 5 NiV proteins. The screening was carried out using 2 softwares, Autodock4 and DOCK6.8. We then selected only those small molecules that were common in the top 150 best scoring poses from both the screenings. Additionally, we imposed a cutoff for geometrical similarity in the pose of the small molecule for increased confidence. Further, we subjected these shortlisted protein-small molecule complexes to MD simulations to assess their stability. The binding free energies were also calculated for these complexes. This hierarchical process of criteria based filtering the best complexes and the assessment of their stability enabled us to have higher confidence in them. To the best of our knowledge, this is novel filtering criteria that can be applied to any virtual screening exercise to get increased confidence.

We predicted 10 such small molecules that fitted into the above mentioned criteria. Additionally, we selected 3 small molecules, one of which binds to the G protein on its ephrin binding interface, the second one binds at the interface of the M protein dimer and third one which binds to the P protein. The 13 ligand bound protein complexes were subjected to triplicate MD simulations (50 ns each) to gauge the stability of the association. In 9 of the complexes, at least one of the trajectories was evaluated to have favorable (negative) binding energy. While the simulations and the energy calculations that follow are not to be construed as indicators of binding strength, they do provide the same general trends and give pointers and/or boost our confidence in the binding efficacy of the ligand-protein complex. Only 3 of the 13 ligands consistently moved away from the original predicted binding pocket even when the simulations were repeated using a different force field. In one other case, though the protein-ligand complex remained conformationally stable throughout the course of the triplicate trajectories, our energy estimates of this interaction were unfavorable (positive energy). In the absence of experimental validation, which we seek to do next, these MD simulations serve as indicators of the viability of the ligands to bind the viral proteins

In all our computational predictions, an independent scoring scheme(s) was used to evaluate results. MD simulations were always carried out in triplicate and sometimes using different force fields. In short, we have taken care to ensure cross validation of our computations to whatever

extent is practically possible. We cannot overemphasize the importance of these computational predictions, especially for swift acting potent viruses such as NiV where mortality rates are high.

An important aspect of the ZINC library is that it consists of already known drugs. In this study, we predicted some of these drugs as inhibitors of the NiV proteins. The advantage of such repurposing is the ease of testing. For instance, we identified Cyclopent-1-ene-1,2-dicarboxylic acid (ZINC04829362) as an inhibitor of the NiV N protein. This compound is a known drug prescribed for antiasthmatic and antipsoriatic among other disorders. In addition to this, we also tried to find overlap between the shortlisted ZINC molecules and those that occur naturally in plants. We did not find any overlap between the two sets of molecules. In future studies, one could include the naturally occurring ones in the virtual screening exercise. If these molecules are shortlisted as putative binders, it would be easy and cheap to test and use them. Similarly, using a combination of small-molecule libraries with different properties or different sources of origin (plant-based, synthetic, extracted from fungi etc) could be included in the screening.



# Chapter 6

## Conclusions and future prospects

This chapter summarizes, concludes and discusses the future prospects of this thesis. The summary and future prospects are arranged into 3 sections, one each for GPCR activation mechanism, Packpred and design of therapeutics against the Nipah virus. This followed by the final section of concluding remarks.

### 6.1 GPCR

#### 6.1.1 GPCR summary

We studied if the activation of G-protein coupled receptors (GPCRs) is a universal process. When the local environment of GPCRs is perturbed by events like ligand binding, they undergo conformational changes leading to their activation. These conformational changes are well studied for only six GPCRs including the  $\beta_2$ -adrenergic receptor and the adenosine  $A_{2A}$  receptor. For a vast majority of the GPCRs, the conformational changes associated with the activation are unknown. In this example, we studied if all the GPCRs undergo activation in a similar fashion, and if not, in how many distinct ways does it happen. To address this, we analyzed 48 and 15 inactive and active state structures of Class A GPCRs respectively, to detect conserved 3D local environments. We represented all these structures as 16 r-groups. r-groups are subsections of 20 standard amino acids grouped by chemical properties. We then define local environments as cliques. To identify similarities in the local environments, we geometrically and chemically compare cliques across all structures. The cliques that have r-groups identically located (in terms of geometry) in at least 70% structures and those that show less variation are considered to be conserved. The variation is quantified using Shannon entropy, with the conserved groups having Shannon entropy less than 1. To gain higher confidence in our results, we performed this analysis in triplicates for both the inactive and the active state structures separately. We identified 18 conserved cliques in both the inactive and the active states, indicating their importance in maintaining the

structure in that particular state. These conserved cliques could also be important for state transition. To identify the cliques that are responsible for the inactive to active state transition, we compared the inactive state conserved regions with the active state. The cliques that are conserved in the inactive state but not in the active state are ones that changed their conformation during activation, and hence are crucial in the state transition. We found 15 out of 18 cliques were either partially or completely disrupted in the active state. The 3 cliques that did not undergo any conformational change, could be important for stabilizing the structure. Furthermore, to identify the cliques that are newly formed in the active state, we compared the active state conserved cliques with the inactive state ones. We found that seven new cliques are formed in the active state. 8 of the 11 cliques that are common to both the states gain new contacts during the activation process, while 3 remain unchanged. We then validated these residues using data from the literature. We could validate ~22% of the 435 reported mutations using our results. Next, we tested the residues that were not validated by the literature data using MD simulations. To summarize, we predicted the residues that are a part of the activation pathway of Class A GPCRs.

### 6.1.2 Suggestions and future directions

In this study, we have exclusively used Class A structures; structures from other classes of GPCRs can also be included in the analysis to check for conserved activation pathway. The number of structures for other classes are limited but their models generated by recent methods like AlphaFold2[90], RosettaFold[91] could be used for the analysis. These structures can either be used to derive the activation pathway or can be used as another dataset that is used to validate the predicted pathway.

The binding site of G-proteins / arrestins on the activated GPCRs is co-localized. Some receptors preferentially bind to G-proteins while some to arrestins. In addition, there are several different types of G-proteins. An algorithm similar to the one designed in this study can be implemented to identify the determinants of association of types of G-proteins or arrestins with various GPCRs.

In this study, we represented the amino acids as r-groups. R-group representation helps to identify the small sections of the amino acids that are necessary for the structure or function of GPCRs. Using multiple definitions of r-groups and then taking a consensus of

the important ones could prove to be a crucial step in gaining higher confidence in the predicted activation pathway. Using multiple definitions may allow us to capture different sets / overlapping sets of important r-groups.

Another promising area for future studies is to modulate the activity of GPCRs by use of cryptic sites. Cryptic sites are short peptides that are embedded in the extracellular membrane and are inactive/dormant. In cases where the constitutively active GPCR is the cause of a disease, we can design such cryptic sites that will partially block the receptor. Similarly, diseases that are caused due to overactivation or underactivation of GPCRs could also be modulated using the cryptic sites[145].

In the field of structural biology, the local environment is accounted for in various ways. It is commonly used to check if a particular region has favorable interactions leading to stability. Such information is often incorporated into a scoring function that allows for assessment of stability of a structural model of a protein. For example, in the rosetta scoring function, local interactions such as Van Der Waals attractive and repulsive forces, backbone and side chain hydrogen bonds, electrostatic interactions, solvation energies and several other terms are included. A weighted sum of these physics based entities forms the final score for a protein structure [146] . Similarly, there are other knowledge based scoring methods that extract various features from the existing protein structures. For instance, another method GOAP [33] uses the data about the angle between planes of heavy atoms that are within a distance cutoff of each other. The distance cutoff ensures that the local interactions are captured in the method. Similarly, multiple such definitions of local environments could be implemented and then a consensus of their results can be considered. Following such workflows will increase the robustness of our method.

GPCRs are known to interact with various lipids like cholesterol. The cliques that are involved in such interactions could be identified using the algorithm that we have developed. This study could be extended to all the membrane embedded proteins as well. Additionally, some of these proteins form oligomeric structures. We could also apply this algorithm to identify the cliques that enable such oligomerization events [147][148].

The idea of modularity of GPCRs can be further explored as a design principle. For instance, using the activation pathway identified from our study, we could design a GPCR that responds to a ligand of our interest. Additionally, the residues that are not a part of the activation pathway can be mutated without affecting the functionality of the receptor.

These residues could be targeted to make GPCRs thermostable/improve their conformational stability, assisting to solve their structures experimentally. Another area where the design principle could be applied is the downstream effect of the GPCRs. Some ligands activate the G-protein pathway while some activate the arresting pathways. If we find the determinants of activation of a specific pathway, we could also alter the pathway a particular ligand activates [149].

All the results in this part of the thesis are purely computational. Experimental validation could be done to confirm the results achieved using this algorithm.

Finally, we think that the algorithm that we have developed is robust with respect to the diversity in a class of proteins (such as GPCR). Hence, this algorithm could be applied to other proteins to identify the determinants that are important for its structure or function, irrespective of the degree of the sequence divergence of its members

## 6.2 Packpred

### 6.2.1 Packpred summary

We developed a tool, Packpred, that predicts if the perturbation in the local environments would be neutral or deleterious for a given protein. Particularly, Packpred accounts for perturbed local environments in the form of mutations. Any given mutation is classified in either of the two prediction categories, neutral or deleterious. To predict the category, Packpred uses a linear combination of three distinct scores, namely, the statistical potential, FADHM substitution matrices and Shannon entropy. The statistical potential calculates the observed by expected ratio of the clique. The rationale for calculating this ratio is that the cliques that are stable are expected to occur more frequently than they would occur by chance. Higher the value of the ratio, better is the stability of the clique. The second score that Packpred uses is the FADHM substitution matrix, which divides a protein into 3 regions (Exposed, intermediate and buried) based on their residue depths. Residue depth is the distance between a residue and its closest bulk solvent molecule. For each depth region, FADHM indicates the probability of substitution of one amino acid to another. The probability is obtained using structure alignments of proteins from the PDB. The third and the final score that Packpred used is the Shannon entropy, which quantifies the evolutionary information from multiple sequence alignments. The linear combination

of these three scores is obtained by training Packpred on a saturation mutagenesis dataset of T4 lysozyme, consisting of ~2000 mutations. We tested Packpred on two different datasets, the CcdB saturation mutagenesis dataset and the Missense3D dataset consisting of ~1500 and ~4000 mutations respectively. The Missense3D dataset is a complex dataset as it consists of mutations derived from ~600 proteins. We compared ourselves with 6 other state-of-the-art methods and showed that Packpred performs at par or better than all these methods on the testing sets.

### 6.2.2 Suggestions and future directions

Although Packpred outperforms all other methods in the Missense3D dataset, its MCC is 0.36, indicating a scope for improvement. The performance can be improved by increasing the training dataset. Although we tried to incorporate more saturation mutagenesis dataset in our datasets, we got little improvement over the existing results with only T4 as a training dataset. One of the reasons for this could be that the saturation dataset is biased in terms of depth of the residues. For instance, proteins usually have more exposed residues than the buried residues, consequently biasing the saturation mutagenesis dataset. This creates bias towards better prediction of the exposed residues than the buried residues. Similarly, a bias is also observed in the frequency of naturally occurring amino acids, which may affect the predictions. Thus, creating training and testing datasets that are well balanced in features including residue depths (as mentioned above), ratio of neutral to deleterious mutations, normalization for frequency of occurrence of amino acids etc. is necessary.

Packpred currently uses a linear combination of three scores for prediction. While with the linear combination, Packpred outperforms most state-of-the-art methods, methods other than linear combinations may result in better performance. One way of implementing it would be to use machine learning based methods or to use higher order equations. However, use of these methods may need a higher number of features that cover other aspects of mutational stability determinants. Another improvement can be to use r-group representation, similar to the GPCR work to derive the statistical potential. The statistical potential of Packpred is derived from 3D structures of proteins as seen in the PDB. It represents the proteins as amino acids. The proteins could be represented as

r-groups, similar to our GPCR work. The r-group representation may help us better understand if interactions of small sections of the amino-acids are more meaningful than the amino acids interactions. Additionally, machine learning methods like neural networks could be trained to predict likelihood of a given clique/ local environment. These modifications may assist in improving the performance of Packpred.

## 6.3 Nipah

### 6.3.1 Nipah summary

We predicted putative small molecules that would bind to various Nipah proteins, that would inhibit their activity. To do this, we docked a library of drug-like small molecules from the ZINC database with all the predicted binding pockets of 5 Nipah proteins. We performed docking using two software, DOCK6 and AutoDock4 for increased confidence. We then selected top 150 docked complexes from both the docking exercises and shortlisted the common small molecules. Further, we shortlisted only those small molecules that had a similar docking pose in the two docking exercises. We then validated these putative binding small molecules using MD simulations. These small molecules could potentially bind and inhibit the activity of Nipah proteins by interfering with their binding to their cognate partners.

### 6.3.2 Suggestions and future directions

In this study, we used a small molecule drug-like library from the ZINC database. Instead of using one type of library, we could use a combination of libraries that exploit various features. For example, we could include the small molecules that occur naturally in plants, or use drugs that are already approved and are used in other diseases (drug repurposing). Including such libraries will help in faster testing and delivery of the drugs in the markets for use.

In addition to using two/multiple different software to perform the virtual screening, we could also use various scoring schemes to score the poses generated by the docking software. Then based on the scores of the various scoring schemes, employ a jury system

to shortlist the molecules.

Methods like AlphaFold2, RosettaFold, etc, provide predicted models of proteins that did not have experimentally solved structures / full length structures. These models could be used to perform docking and MD simulations, thus allowing us to use the entire proteome as drug targets.

To the best of our knowledge, the workflow that we have developed to get putative binding small molecules is novel. Performing these steps has helped us gain more confidence in our results. We believe that applying this workflow to other docking / virtual screening exercises would assist in shortlisting candidates with better confidence.

We have attempted to validate the binding of a small molecule inhibitor to its target protein by running 50ns MD simulations. In cases where the small molecule does not leave the binding pocket, running simulations for a longer duration would be helpful to better analyze stability. Additionally, in such cases, enhanced simulations such as accelerated MDs may be performed that will allow the ligands to move out of the binding pockets faster. Apart from this, running simulations with different starting conformations (of both ligand and protein) can also be tried.

We have predicted small molecule binding pockets using a program called DEPTH [42]. DEPTH predicts the binding pockets based on how accessible the surface is to the bulk solvent. DEPTH does not predict cryptic binding pockets. Cryptic binding pockets are the sites that are not detectable on the unbound protein, but become obvious on binding of some drug/biomolecule. Programs such as PocketMiner [150] could be used to predict such sites that can subsequently be used for docking. Additionally, the newly developed AI/ML based methods could also be used in the virtual screening exercises. For instance ML based scoring schemes such as AKScore [151], DeepDTA [152] could be used to score the protein ligand complexes to identify the best binders [153]

Finally, in this study, we have used MM/PBSA to calculate binding energies. The binding energies are often calculated by ignoring the water molecules. Even the waters that mediate the ligand - protein interactions or are in close vicinity of the interaction interface are ignored. The effect of such water molecules could be calculated using other methods such as Nwat-MMGBSA. Binding energies that are calculated by including the water molecules are shown to better correlate with the experimentally determined values [154].

## 6.4 Concluding remarks

In this thesis, we have attempted to study the role of local environments in the structure and function of proteins. Towards this, we have developed a novel algorithm to identify/predict the local environments that are important for structure and/or function of the proteins using GPCRs as an example. This algorithm takes into account geometric and chemical similarities to identify conserved local environments. Based on the findings, we also designed mutations (consequently modified local environments) for validation. We believe that this algorithm could be applied to any other protein family to identify the important local environments. The important local environments could then be used to design proteins of our interest.

Further, we wanted to identify the contribution/importance of each local environment in proteins. Towards this, we wanted to gauge the relative stability of local environments that occur in proteins naturally. We studied the observed by expected ratio of all naturally occurring local environments in proteins (as seen in the PDB database). The observed frequency was derived from the PDB database while expected was calculated to indicate the local environment that occurs purely by chance. We incorporated this information in the form of Packpred. Packpred predicts the effect of mutation on a protein.

Finally, we performed docking and MD simulations to predict putative inhibitory small molecules against the Nipah virus. In this study, we designed a new workflow to predict putative inhibitory small molecules with higher confidence.

To summarize, this thesis shows new insights into how the local environments play a role in structure and function of protein and shows promise in using these findings in the design of proteins that have desired characteristics/properties.



# Chapter 7

## Appendix

This chapter describes various steps/analyses that we performed to reach the final versions of algorithms presented in chapter 3 and 4. The first section of this chapter describes the algorithmic modifications for detecting conserved 3D motifs from the GPCR work(chapter 3). The second section deals with the preliminary analysis that we performed on the T4 saturation lysozyme dataset for Packpred(chapter 4).

### 7.1 Progression of algorithm development to detect 3D conserved local environments in GPCRs

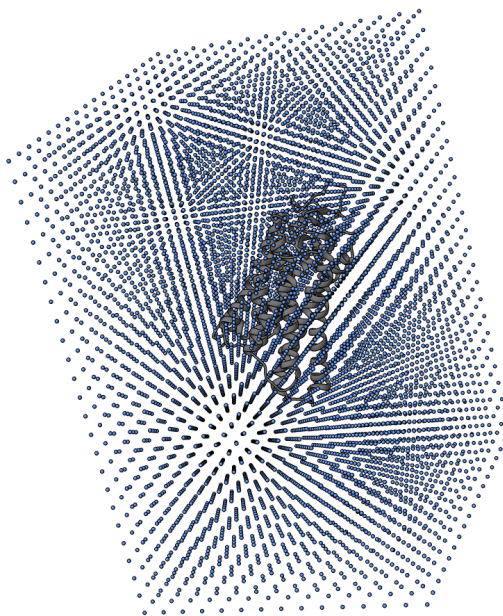
The algorithm that we have reported in this chapter is the version where we decided to report the findings. Since the beginning of the project, this algorithm has been undergoing constant modifications and refinements. This section will briefly talk about the versions of major changes of the algorithm that we tested and will also discuss other ideas that were partially explored.

#### 7.1.1 Versions of algorithm development and refinement

1. We selected all Class A GPCR structures from the GPCRdb. We represented them as amino acids. We defined the local environment as all the residues that lie within a distance cutoff. We defined 3 variants of local environment at distance cutoffs of 5 Å, 6 Å, and 7 Å. Then we performed the local structure superimposition using the software CLICK[3, 4] using Ca as representative atoms. We selected a reference structure against which all structures would be superimposed and compared. For each distance cutoff, we then identified the chemical similarities from the local superimpositions based on the a) Sum of BLOSUM62 scores of the aligned pairs b) Residues that do not align with a partner are penalized with a score of -1. All the reference amino acids that have a positive score and found a partner in the alignment with at least 70% of the dataset are considered conserved.

Findings from this version

- 1.1. Because we took the entire dataset of Class A GPCRs, the over represented GPCRs were biasing our results.
  - 1.2. The use of a reference structure was also affecting the results as the BLOSUM62 scoring etc was with respect to it.
  - 1.3. The penalty score was a heuristic
  - 1.4. Smaller sections of the amino acids could contribute to the stability of the clique
2. We made 4 modifications to the algorithm, first by taking only the best resolution structure per GPCR. The second modification was to use an external grid instead of a reference structure(Figure 1) and then define local environments around grid points. The third modification was to design a different scoring scheme for match and mismatch. The final modification was to use r-group representation instead of amino acid.



*Figure 1: A grid (coloured in blue points) is generated around a GPCR. The grid points are used to define local environments.*

Findings from this version:

- 2.1 Use of external grid now generates dependencies on the density of the grid. As the density of the grid changes, the results would change.
- 2.2 We had defined 3 different r-groups for aromatic residues. Even though they had comparable

properties, these matches were scored unfavorably.

3. In this version we decided to not use the external grid and instead use a reference structure, the approach we had used in version1. Because of 2.2, we clubbed some r-groups together, for instance r14, r15 and r16 (all containing aromatic rings) would be treated as a single r-group. We tried multiple such groupings to check if we get higher conservations.

Findings from this version:

3.1. The r-group regroupings were made arbitrarily

4. In this version, we decided to not regroup r-groups. Instead score the matched/aligned pairs based on the identity of the r-groups from the reference structure. To eliminate the dependency on the reference structure, we used 3 different reference structures and took their consensus.

Findings from this version

4.1. CLICK was not able to correctly align the local environments. The local environments had higher representation of groups like r2, r8, r12 and as a result, the alignments were biased towards these groups.

5. In this version, we used Kersleys 3D least squares fit algorithm instead of CLICK to perform the local alignments. We also noticed that performing a second round of alignments might lead to an improvement in identifying similarities.
6. So now, we perform two rounds of local alignments. But the problem of matching the r-group identity to that of the reference group was still not resolved.
7. Finally we used Shannon Entropy to evaluate the chemical conservation.

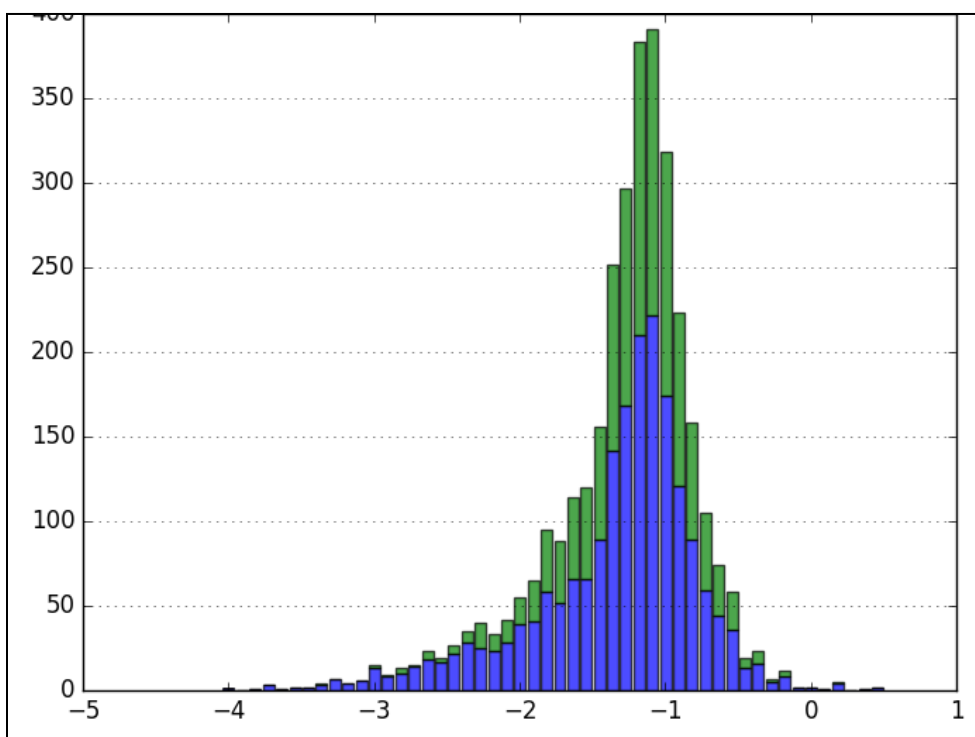
## 7.2 Preliminary analysis that we performed on the data presented chapter 4 Packpred

1. We checked if individual scores, namely, statistical potential, FADHM and SE are able to tell apart the deleterious from the neutral from the T4 lysozyme training set, when used independently and not as a combination. We plotted histograms of the three scores for disease and neutral mutations to check if there is a clear score cutoff that distinguishes the deleterious from the neutral. We observed that both the categories of mutations had overlapping scores for all the three scores and there was no clear separation between the scores of the two classes. We did not observe any trend/cutoff in the scores that would clearly distinguish the two categories of effect

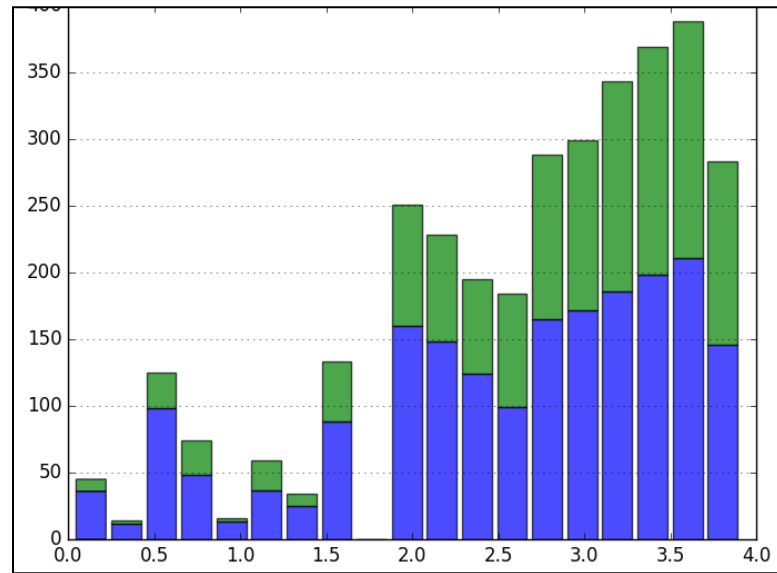
of mutations (Figure2 - A, B, C).

Next, we wanted to check the impact of a parameter, residue depth, on the effect that a mutation has on the structure and function of a protein. So we segregated the mutations based on their residue depths into the three depth bins: exposed, intermediate and buried. Thus, we had a total of 9 categories, 3 depth levels for each of the 3 scores. In all the 9 categories, we observed that the scores for both types of mutations can take up similar values and no clear distinction cutoff was observed that separates the deleterious from the neutral mutations (Figure 3). Thus, we realized that the effect of the mutated environment cannot be quantified by either of the three scores alone.

2A



2B



2C

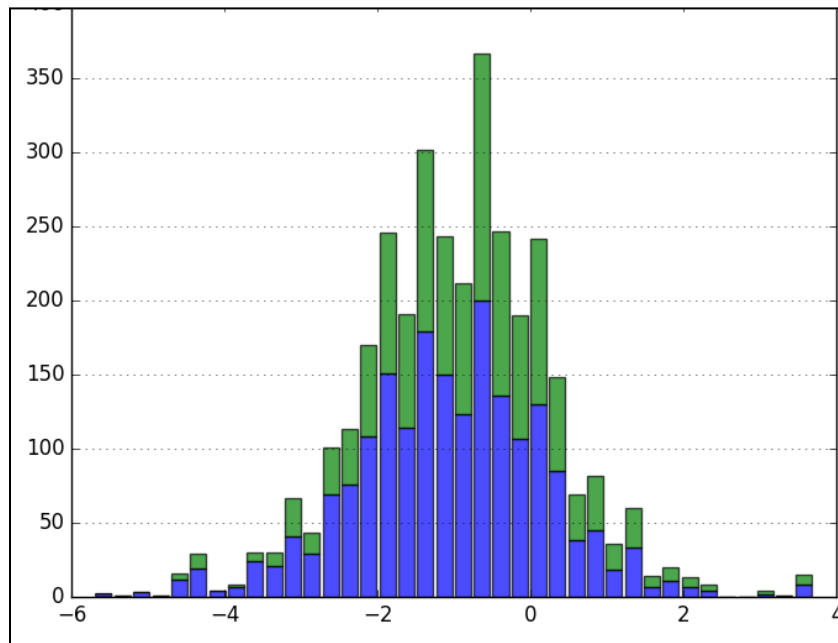


Figure 2: Histograms for (A)PP(Statistical potential score), (B)FADHM and (C)SE. Frequency of deleterious mutations is coloured in indigo while that of the neutral is coloured in green. All of the plots show the presence of both types of mutations across different scores. There is no clear demarcation between the neutral and deleterious.

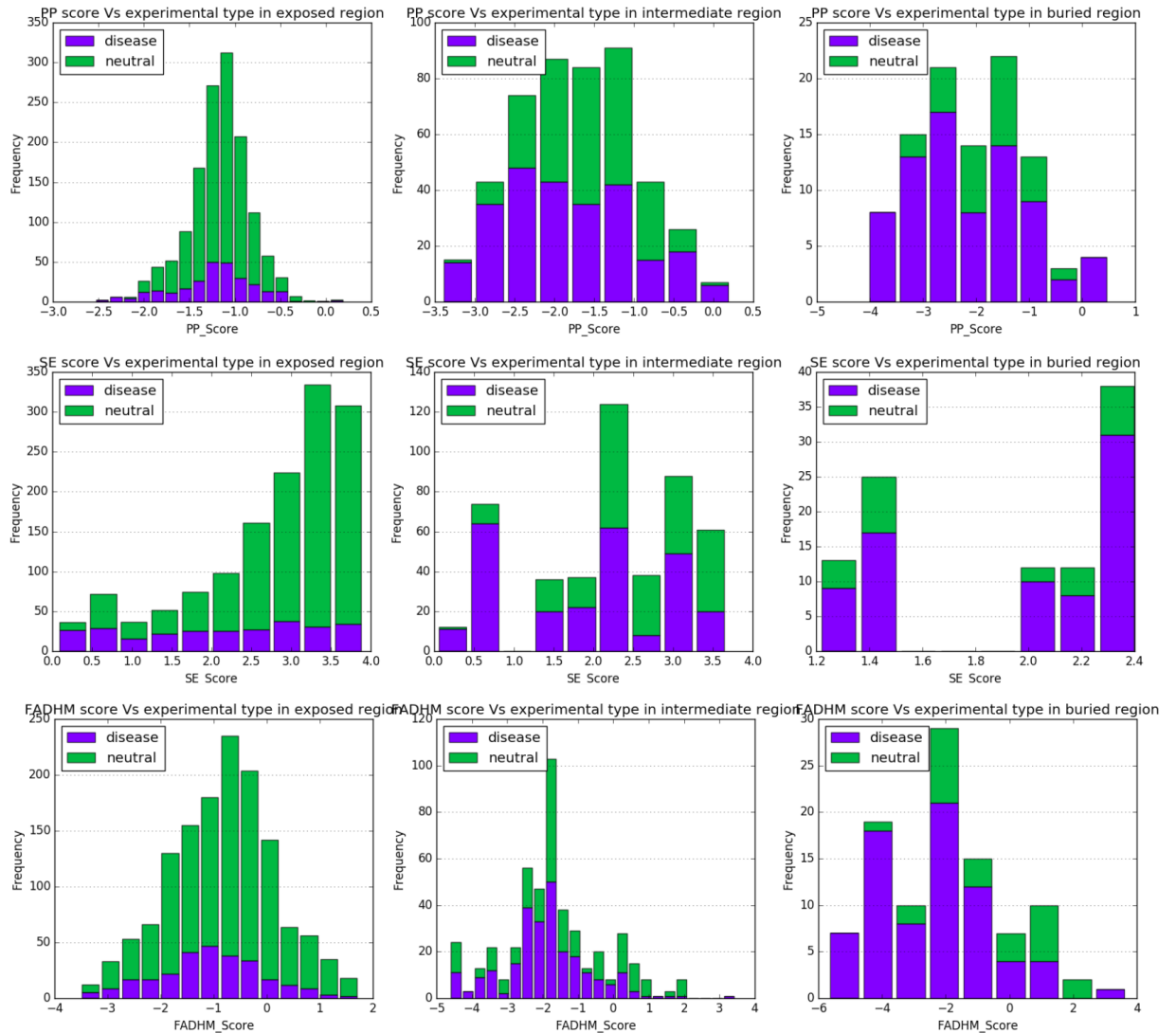


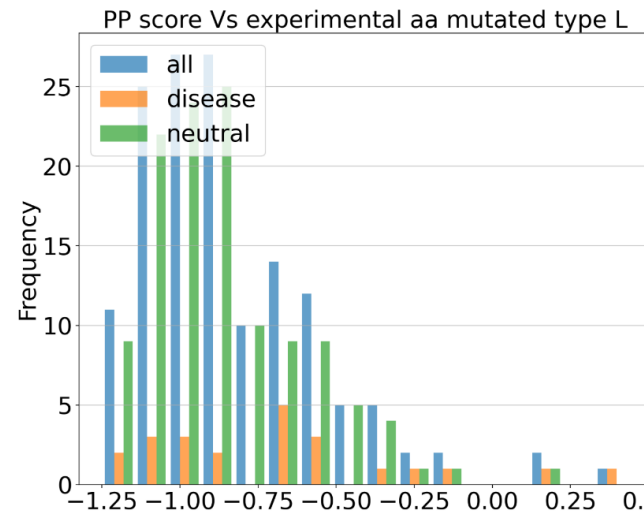
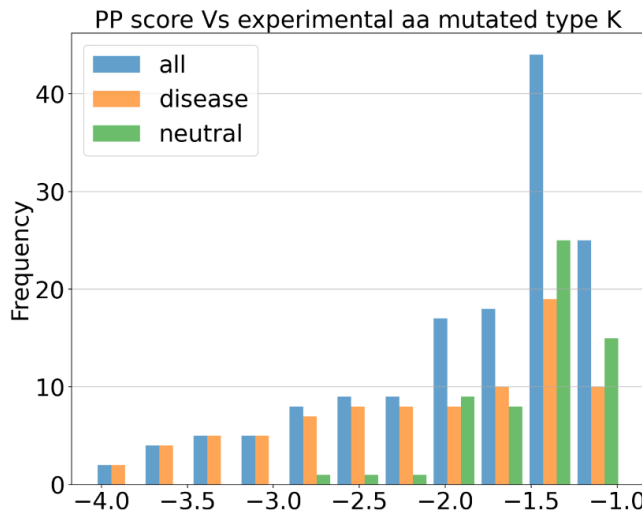
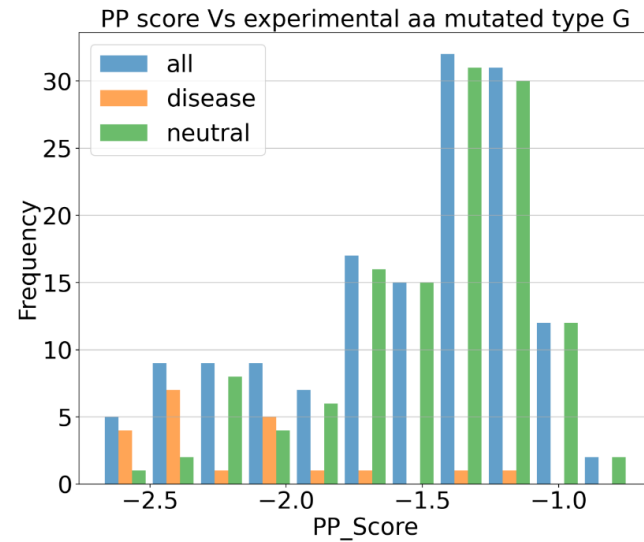
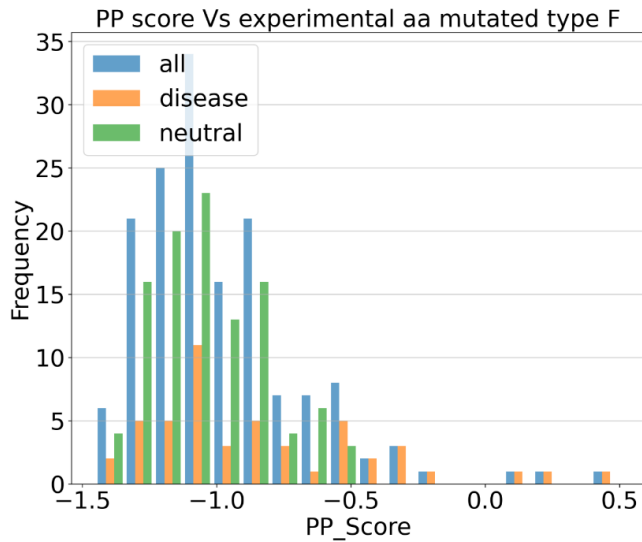
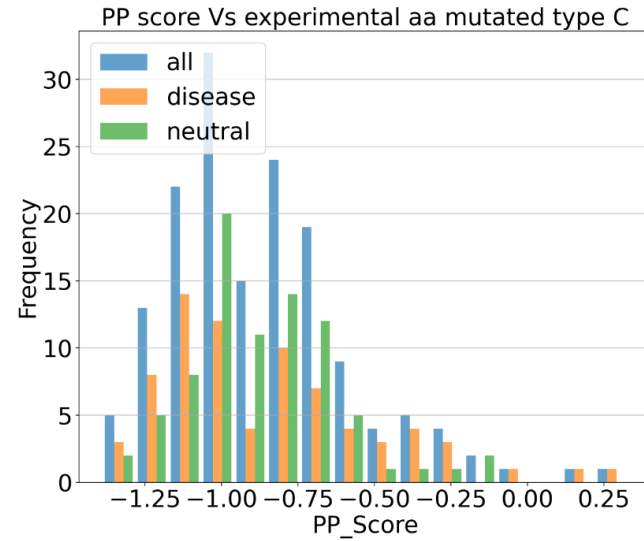
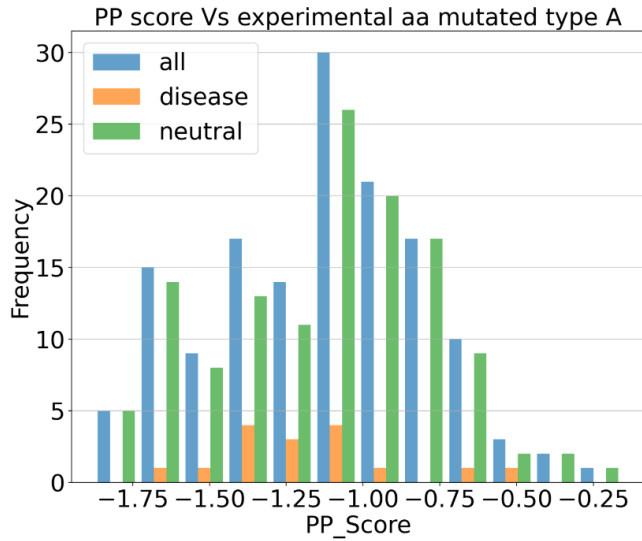
Figure 3: Histograms for PP(Statistical potential score), FADHM and SE across three depth bins, namely, exposed, intermediate and buried. Frequency of deleterious mutations is coloured in indigo while that of the neutral is coloured in green. All of the plots show the presence of both types of mutations across different scores. There is no clear demarcation between the neutral and deleterious types of mutations.

- Next, we wanted to check the effect of mutation when the data is categorized according to the type of amino acid that it is mutated to. We categorized the three scores independently as per the mutated amino acid to observe the score profile(Figure 4, 5 and 6). Since the T4 lysozyme

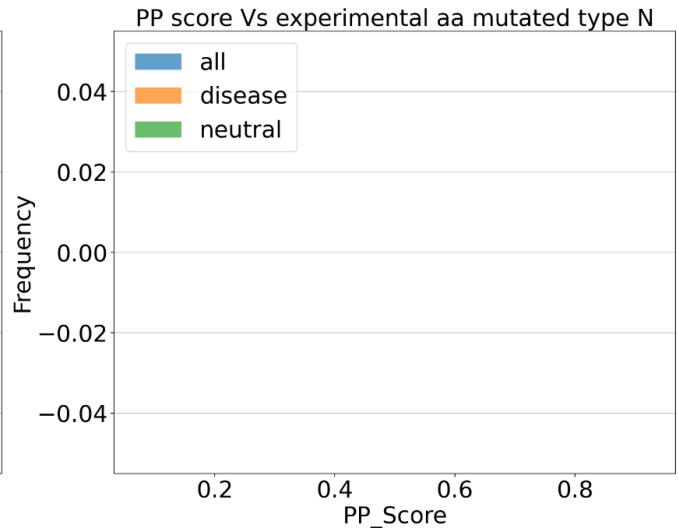
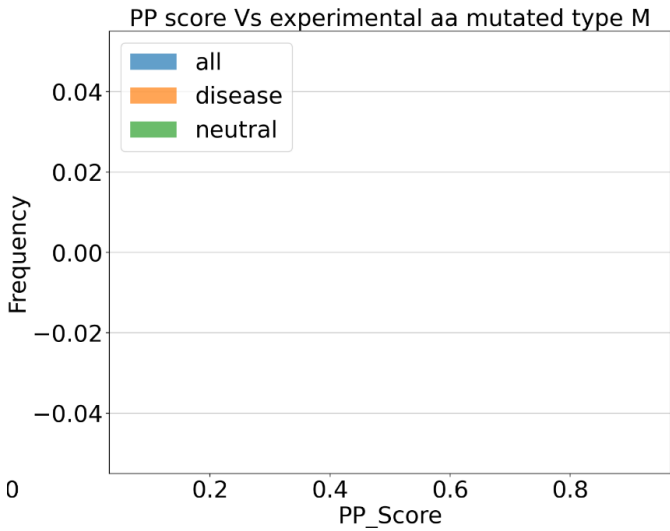
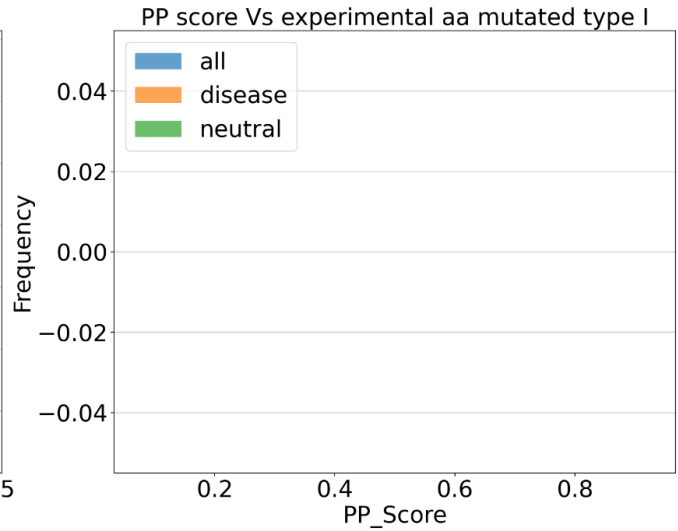
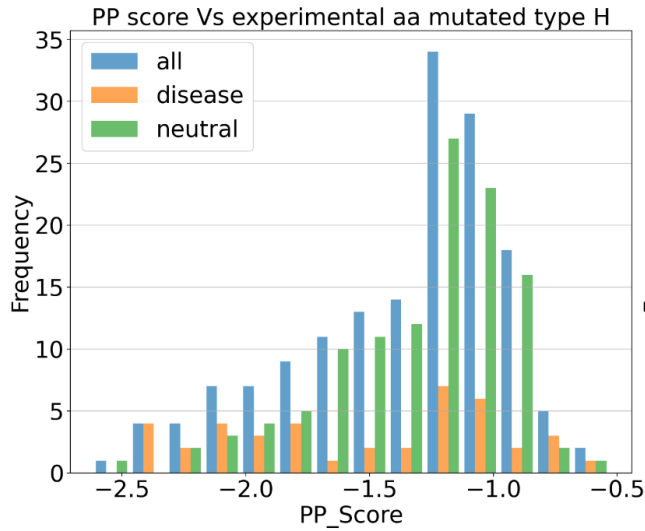
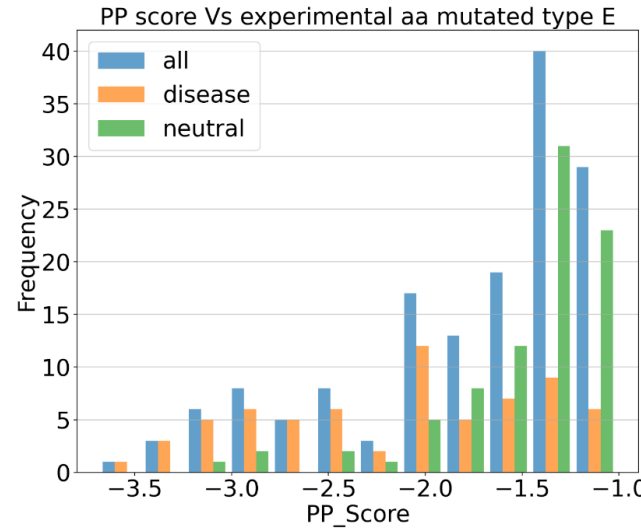
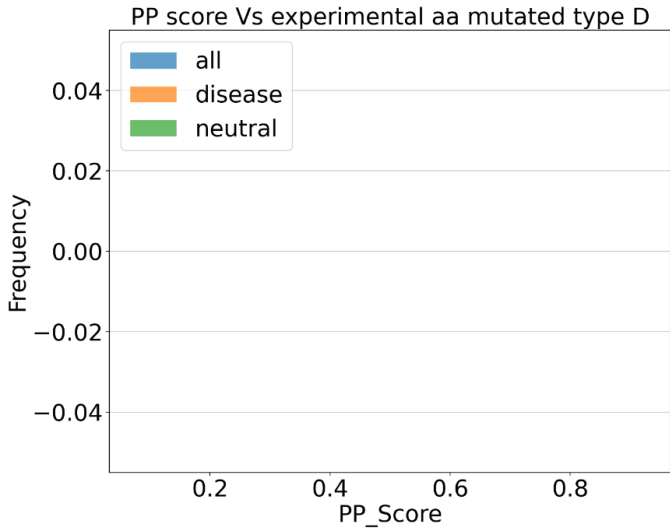
saturation mutagenesis dataset has all amino acids mutated to only 13 other types of amino acids, we did not have data for 7 amino acids. Nevertheless, we analyzed the score distributions of the remaining 13 amino acids. Similar to the observations in Figure 2 and 3, here also we observed statistical potential scores for neutral and deleterious mutations taking up similar score values (Figure 4). Interestingly, we noticed that when the residues were mutated to K, only deleterious mutations occupied scores in the bins having a score less than -3.00. But we also observed deleterious mutations that had a score higher than -3.00. For the remaining 12 amino acids, we did not notice any clear differences between the scores of the neutral and deleterious mutations.

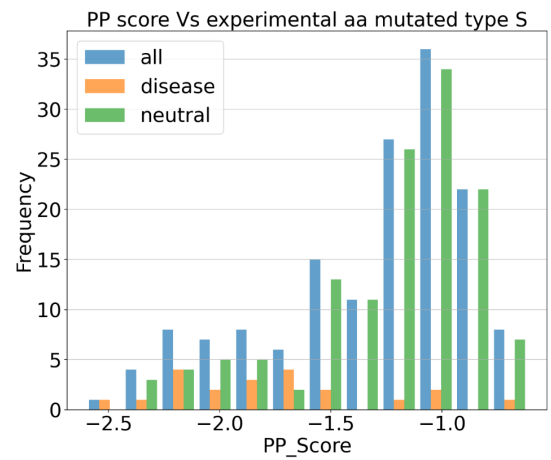
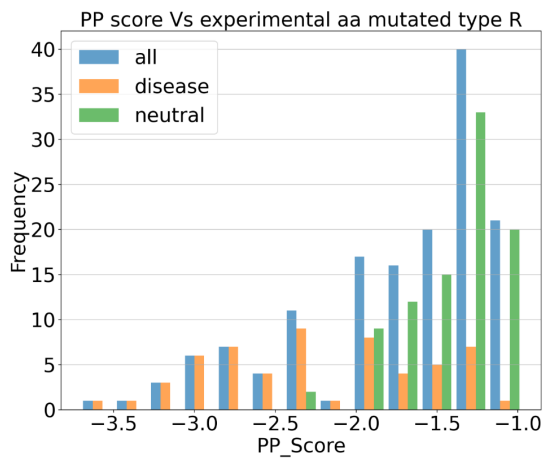
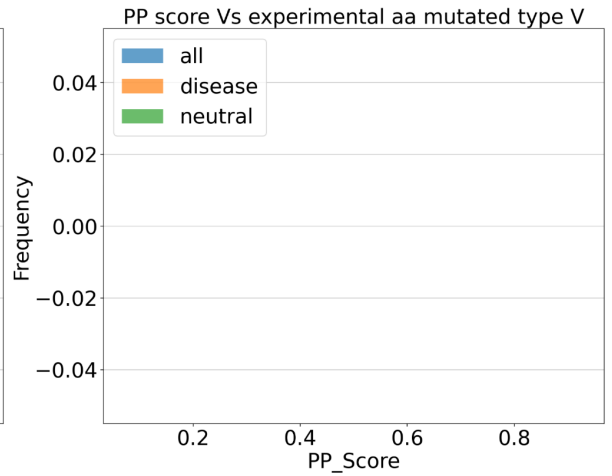
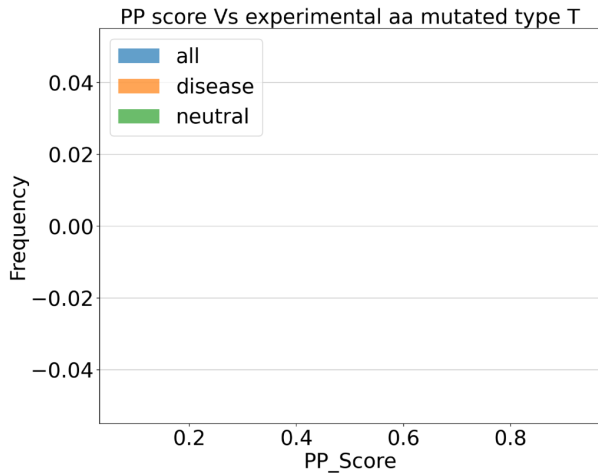
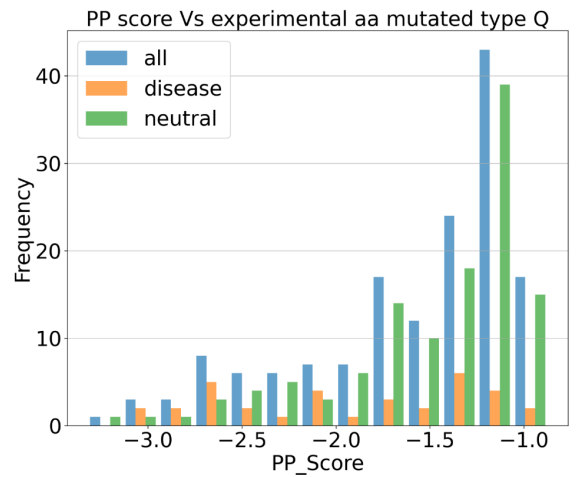
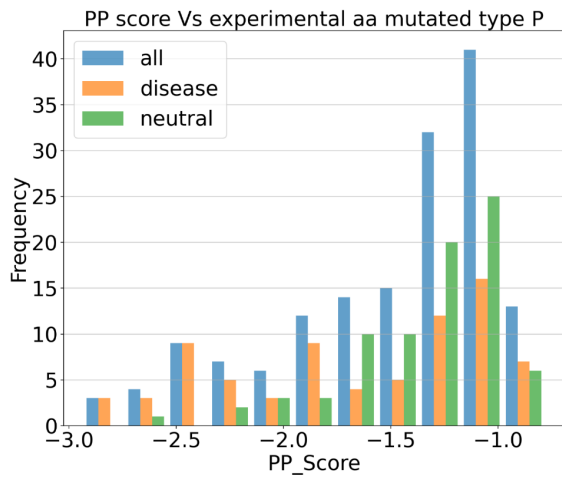
For the Shannon entropy, we observed that for amino acids I, L, H, and Q, the number of deleterious mutations reduced as the score approached from 3 to 4 (Figure 5). Since the most variable position gets a score  $\sim 4$ , these findings were indicative of Shannon entropy being a meaningful score in our study. Additionally, for amino acids C, K, and P we observed a higher number of deleterious mutations between scores 0 to 1, indicative of highly conserved residues.

Similarly, for FADHM scores, we observed that amino acids C, E, K, P had a majority of deleterious mutations with low scores (the range varies with each amino acid) while residues Y, S, R, and Q had only neutral mutations that scored in the positive range (Figure 6).









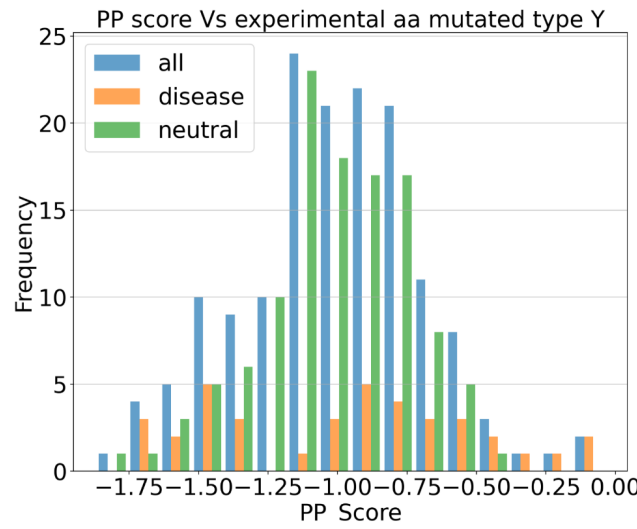
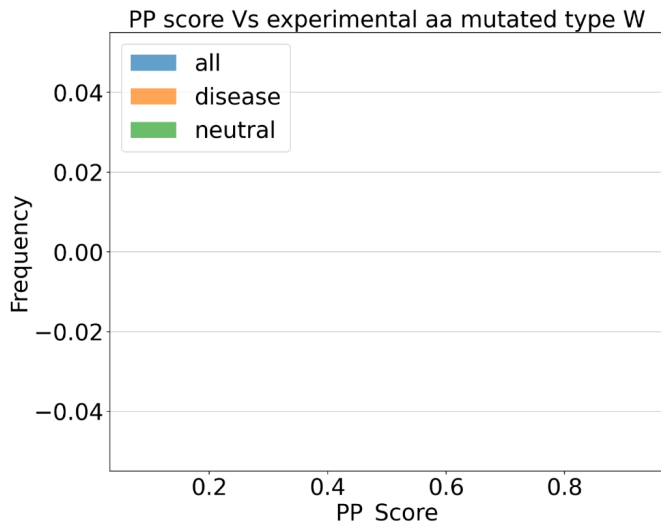
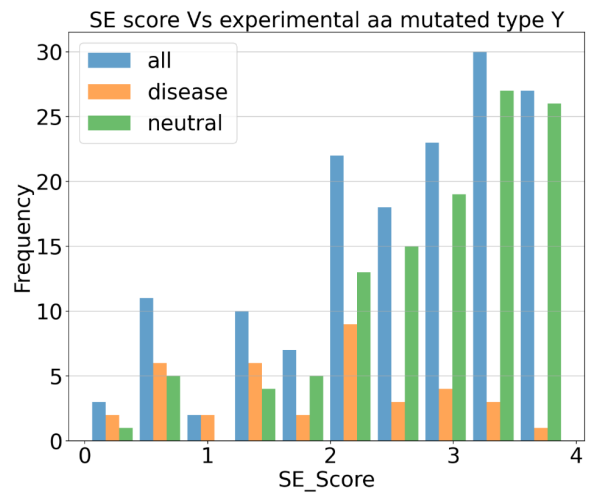
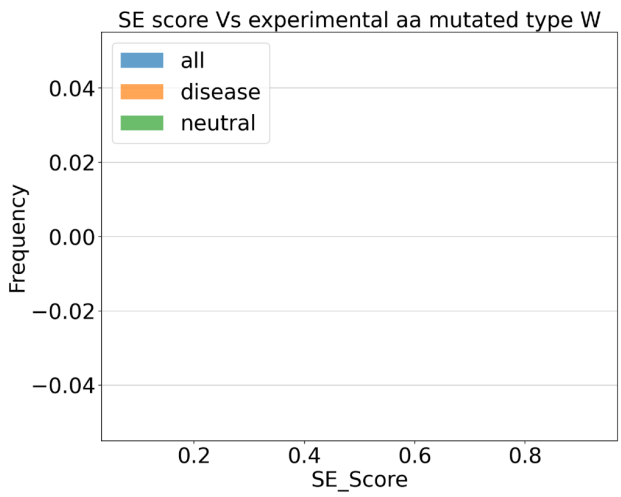
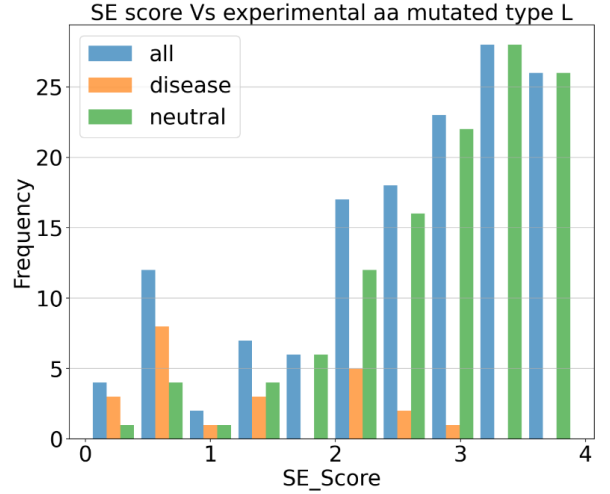
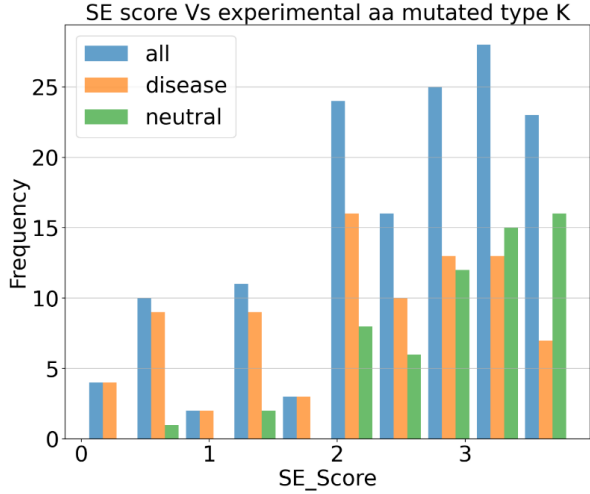
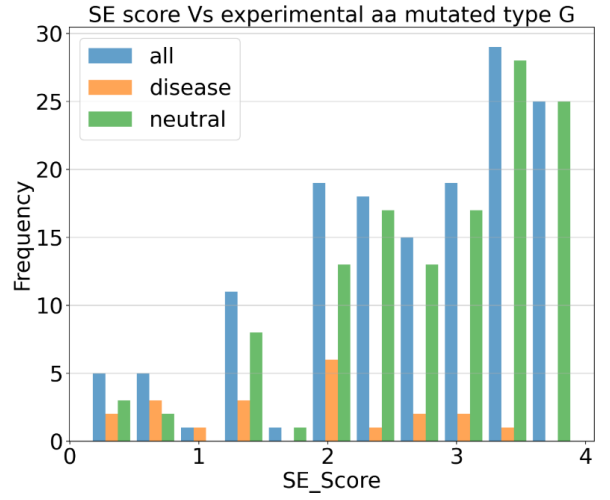
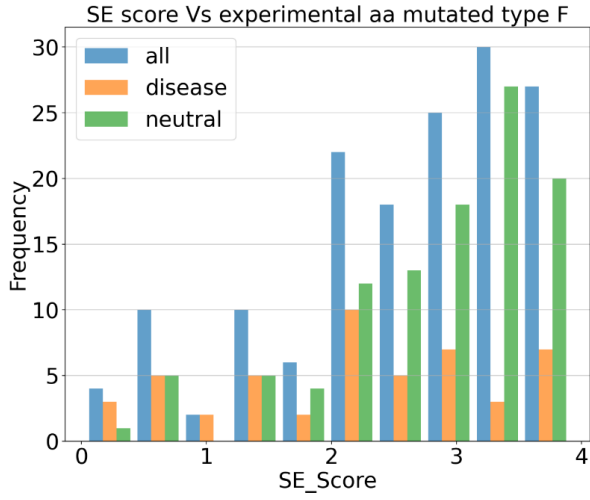
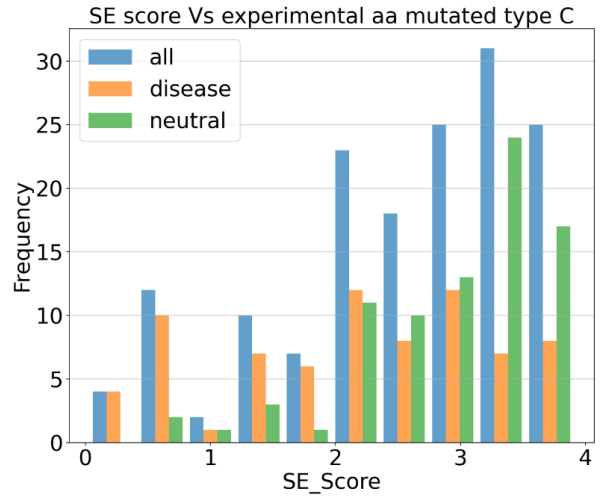
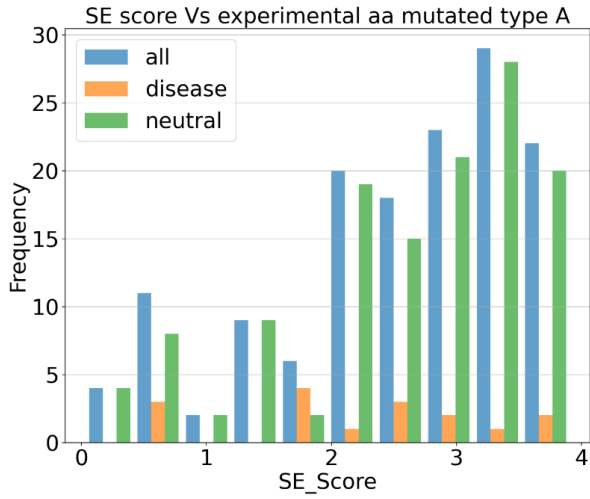
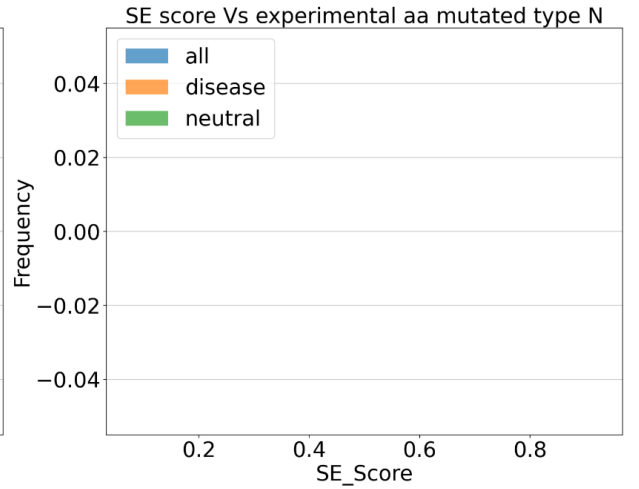
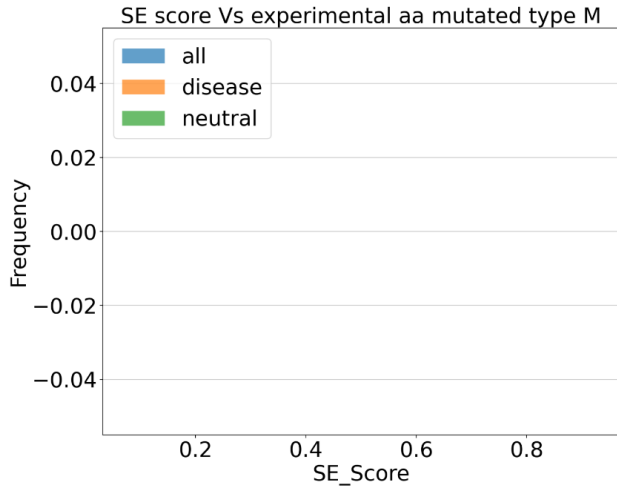
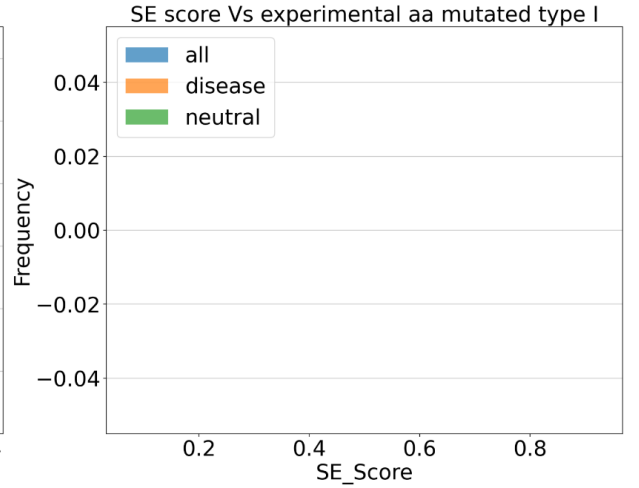
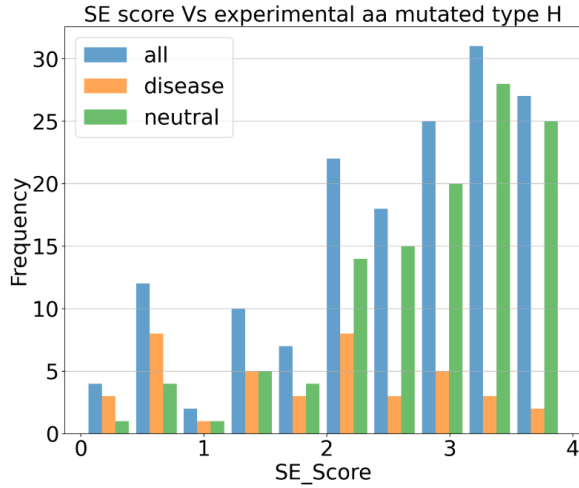
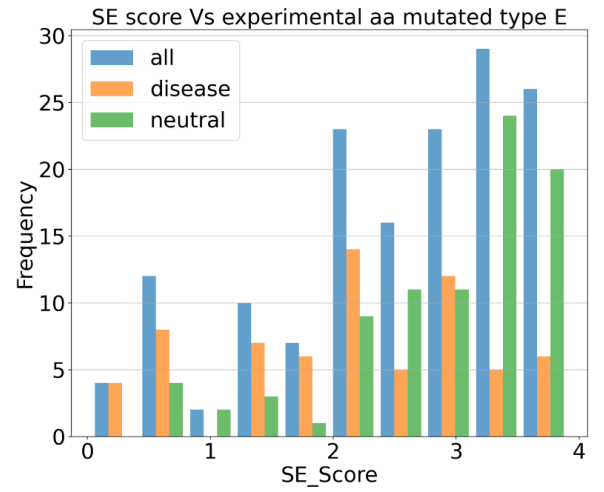
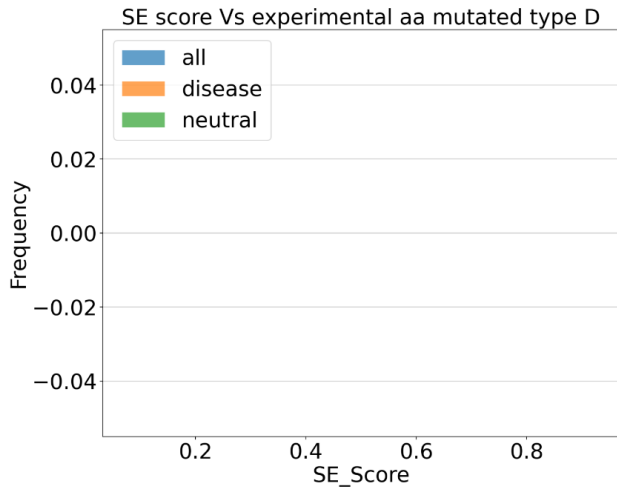
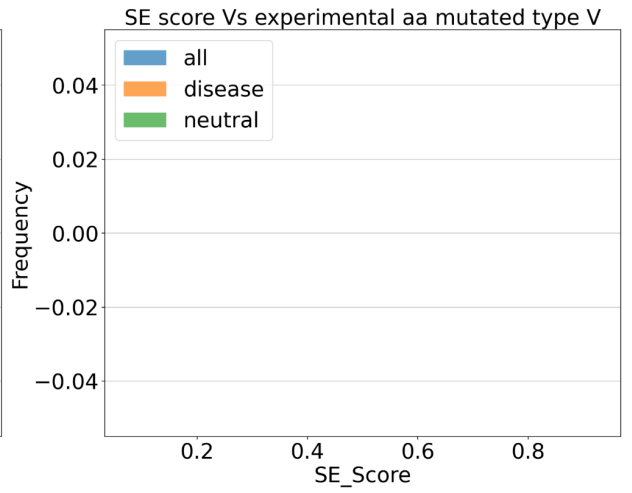
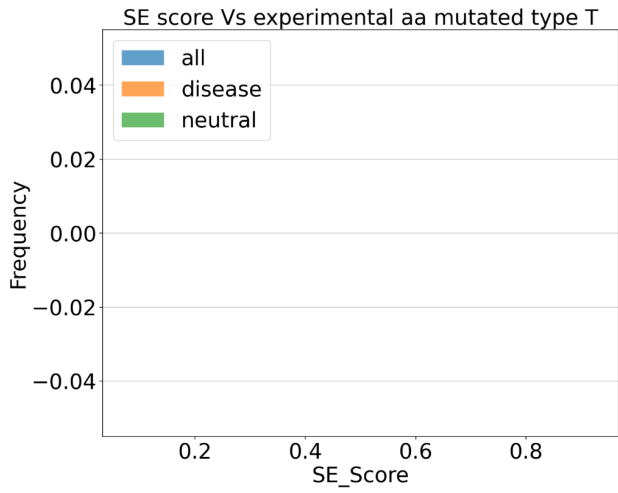
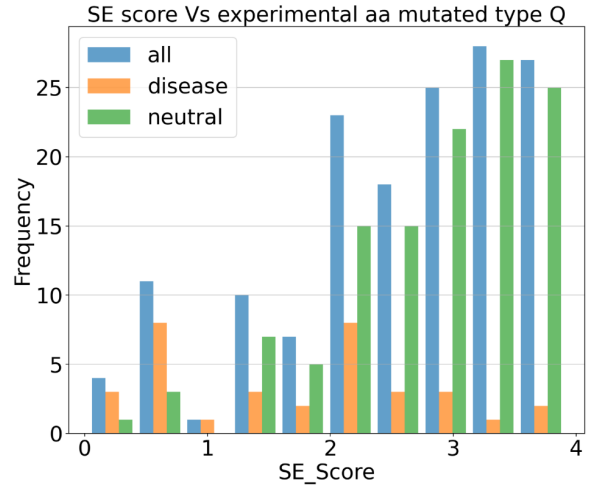
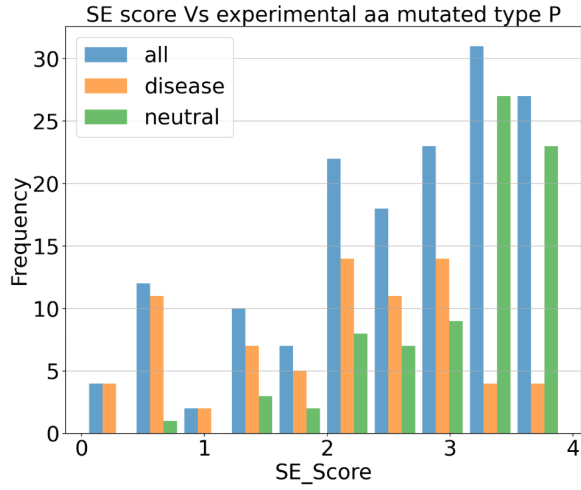


Figure 4: Histogram of statistical potential score categorized according to the type of mutated amino acid. Deleterious are coloured in orange, neutral in green and deleterious combined with neutral is colored in blue. Blank plots indicate no wildtype residue was mutated to this particular amino acid.









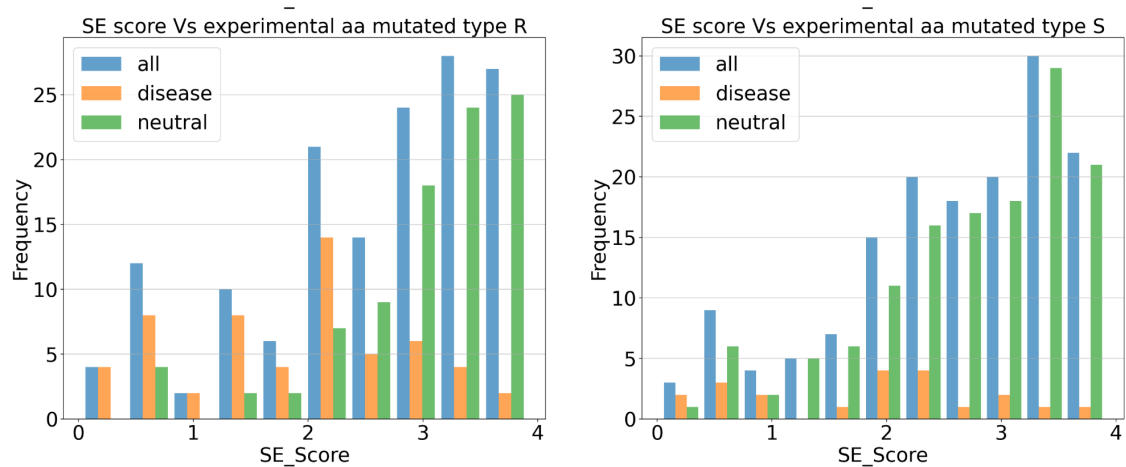
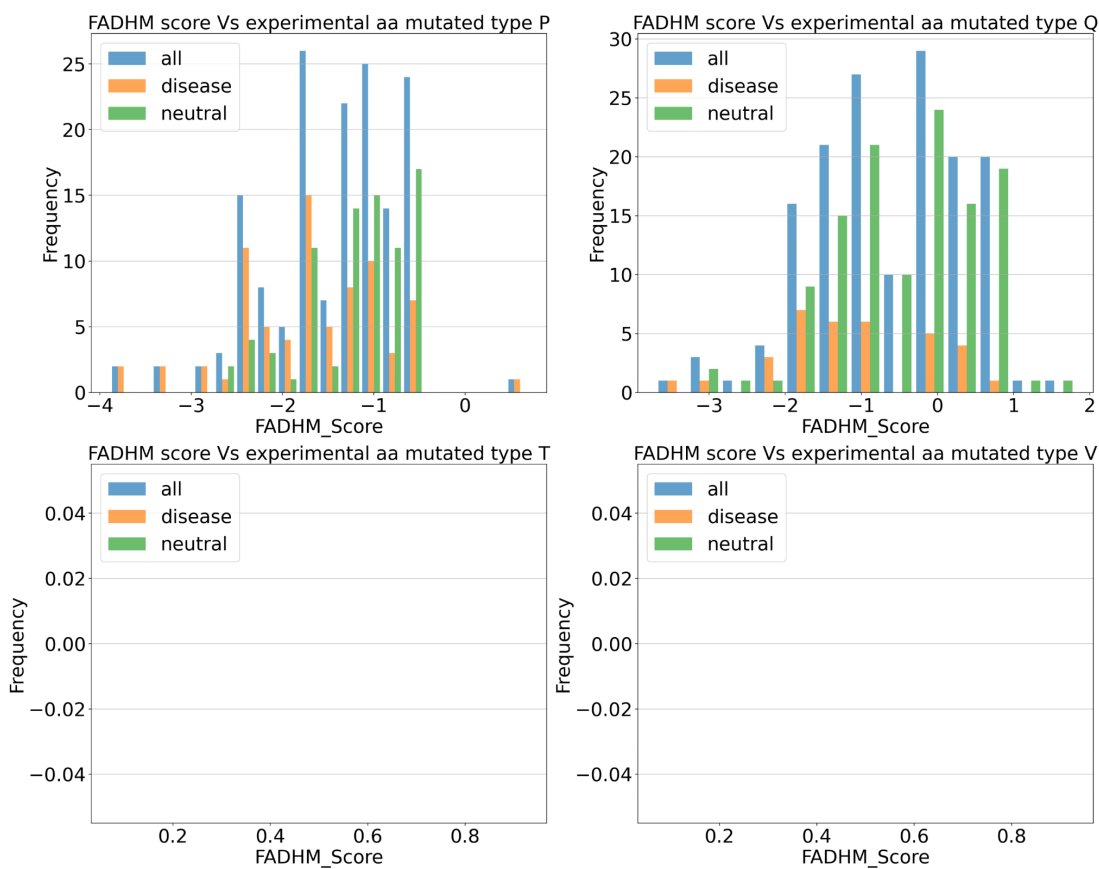
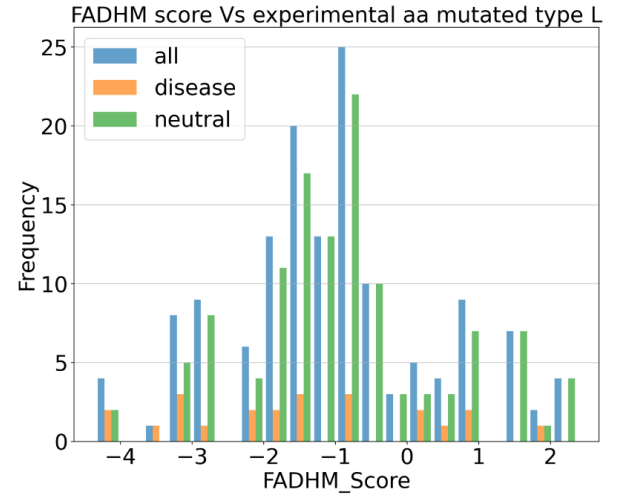
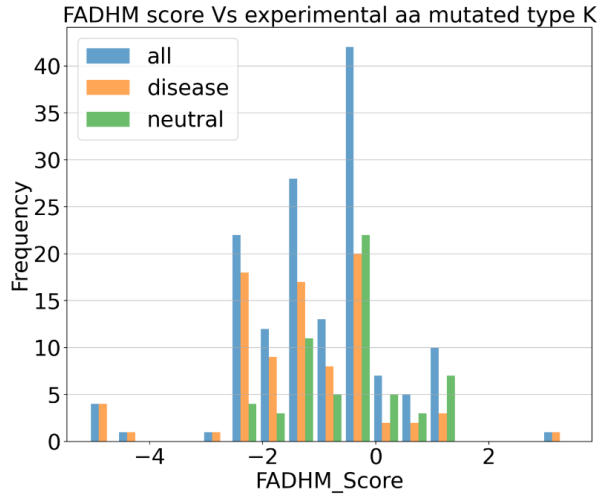
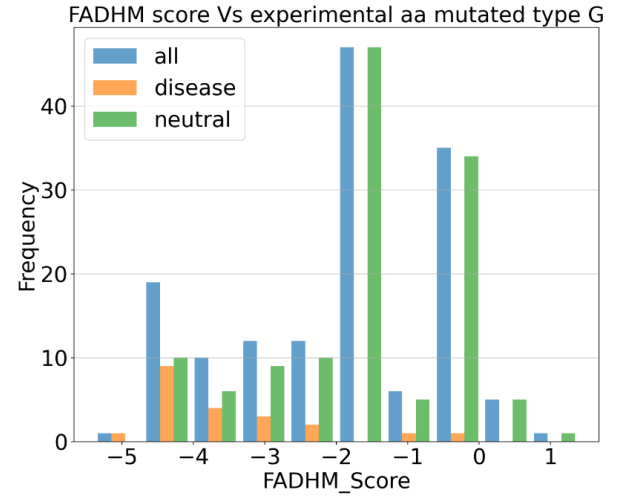
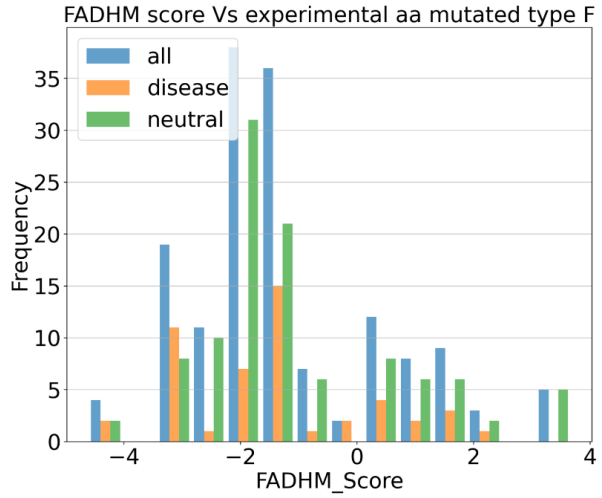
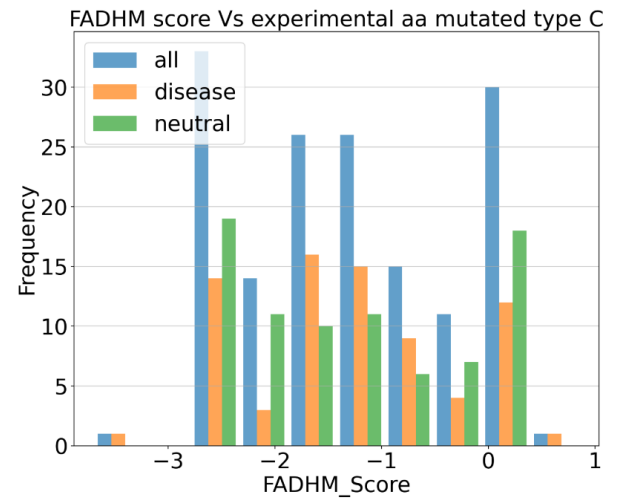
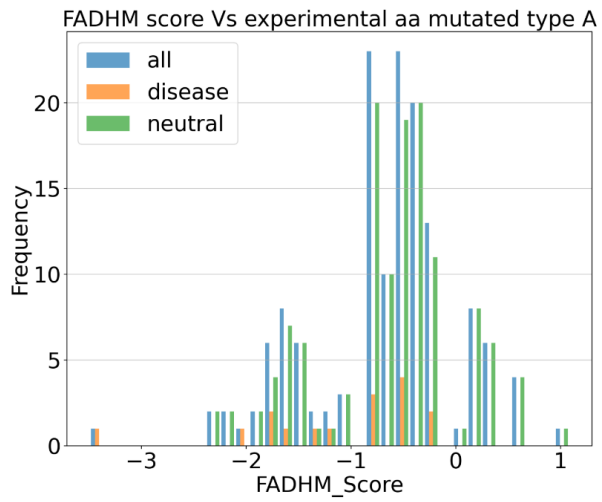
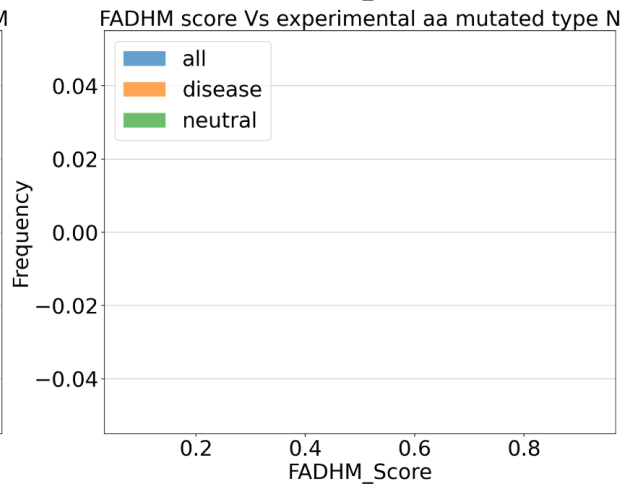
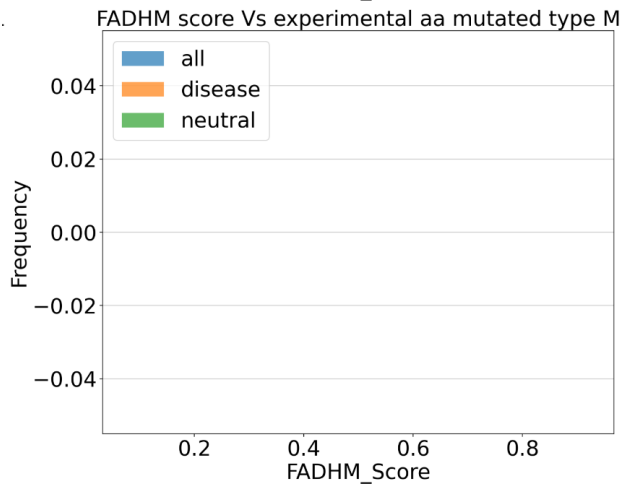
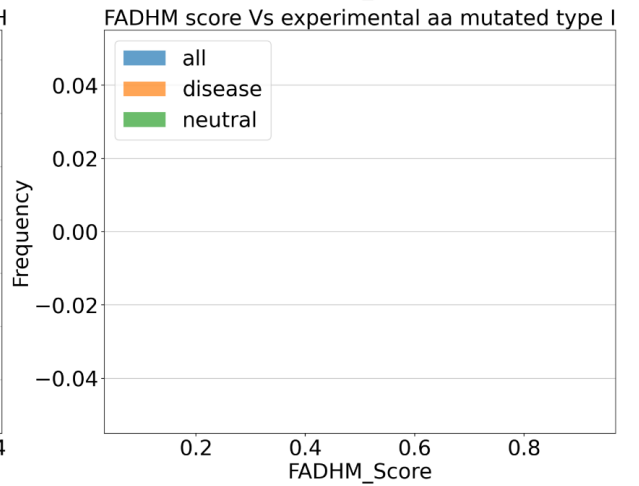
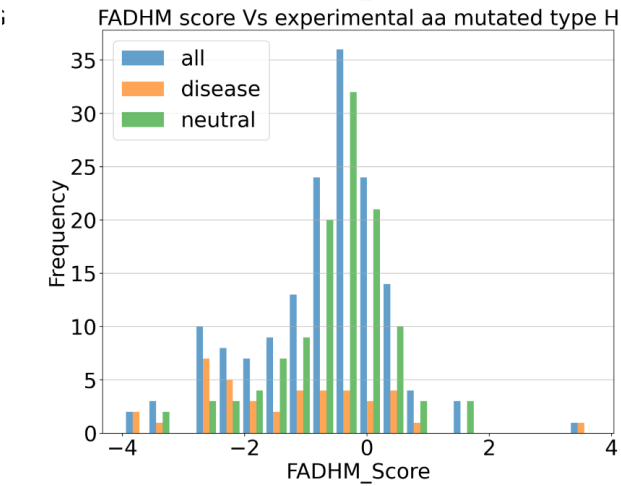
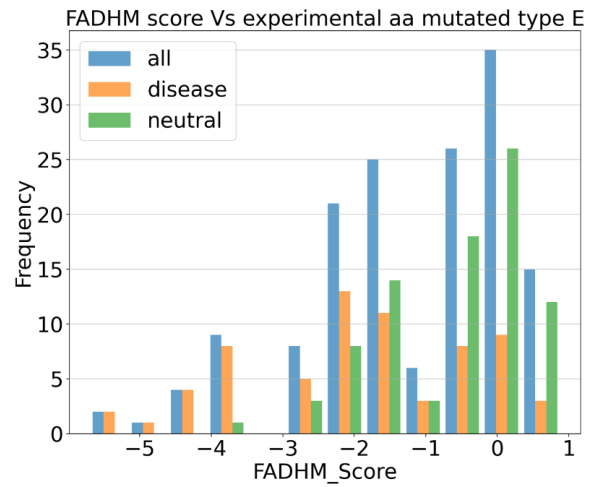
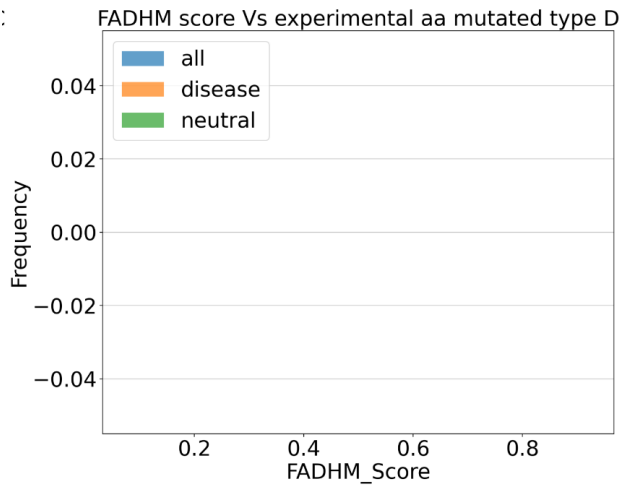


Figure 5: Histogram of SE categorized according to the type of mutated amino acid. Deleterious are coloured in orange, neutral in green and deleterious combined with neutral is colored in blue. Blank plots indicate no wildtype residue was mutated to this particular amino acid.









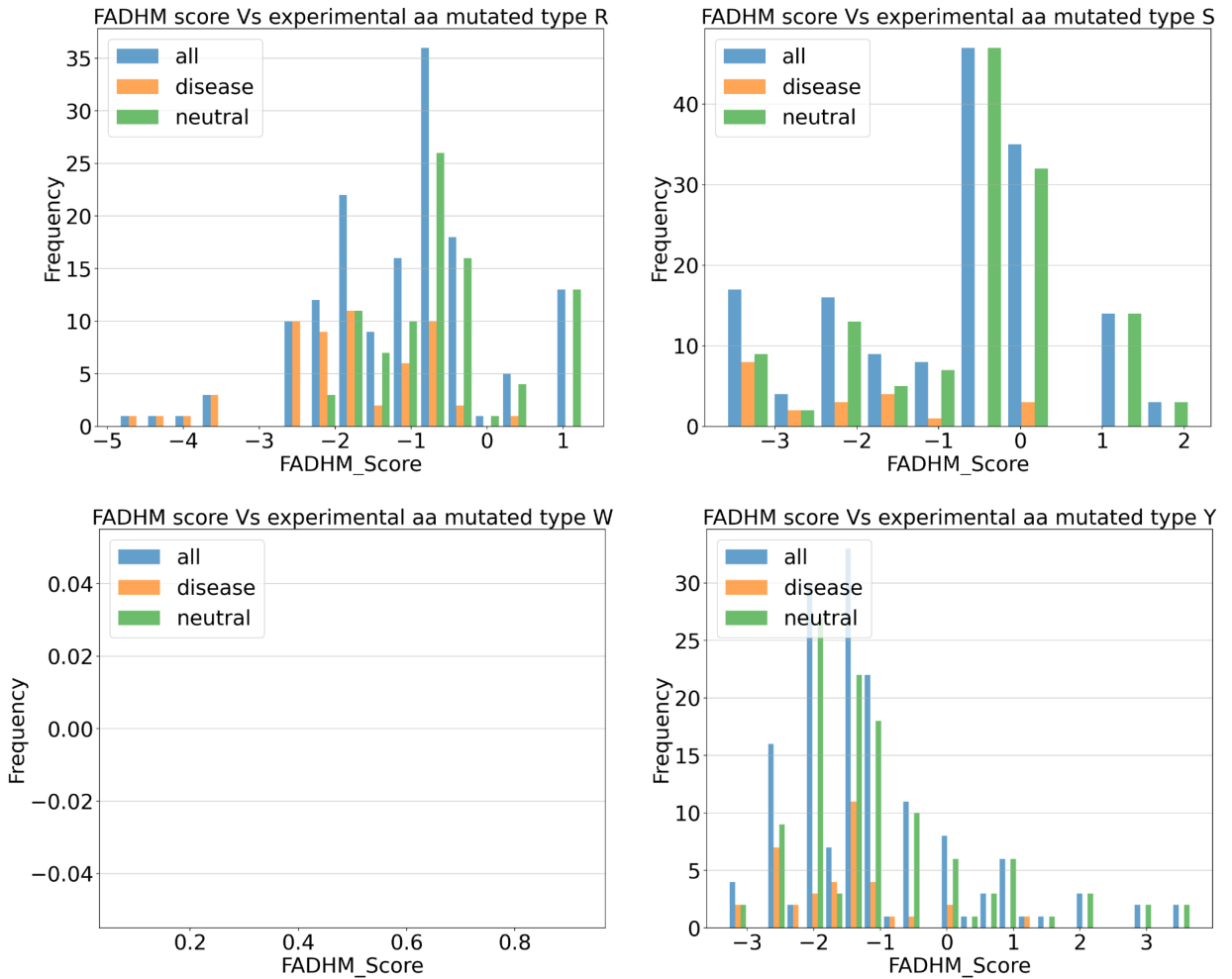
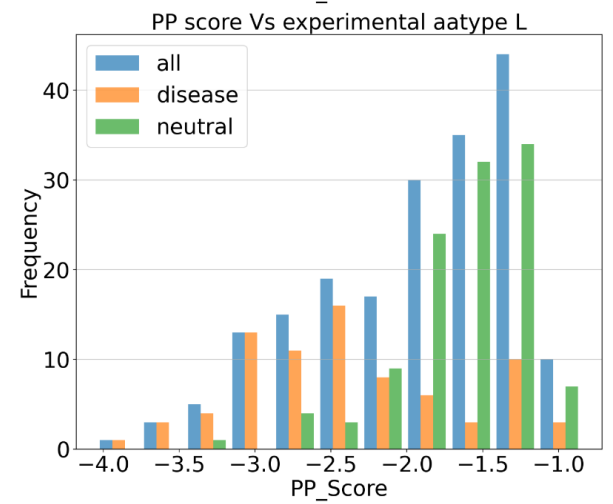
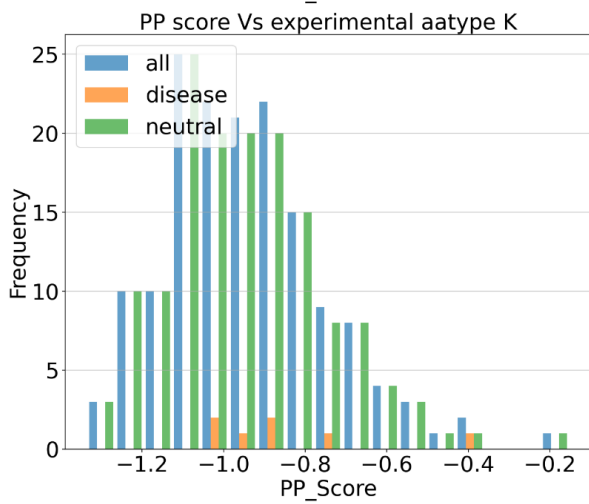
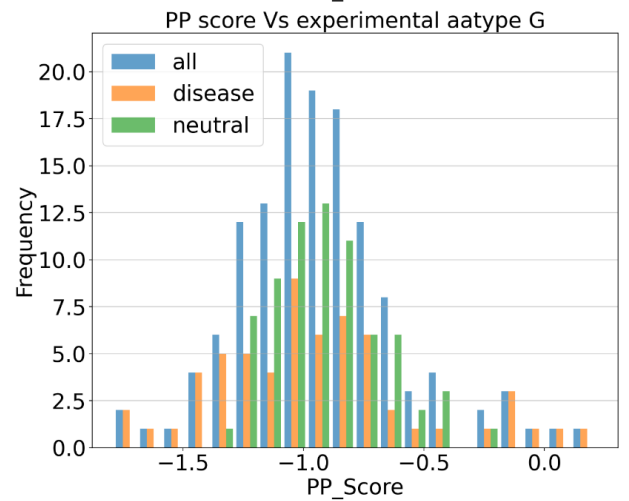
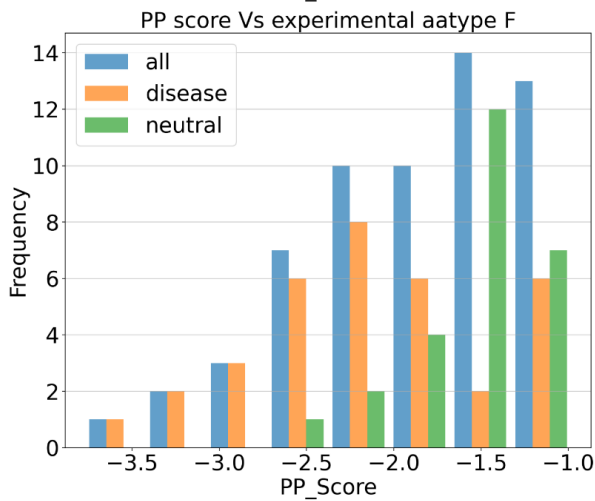
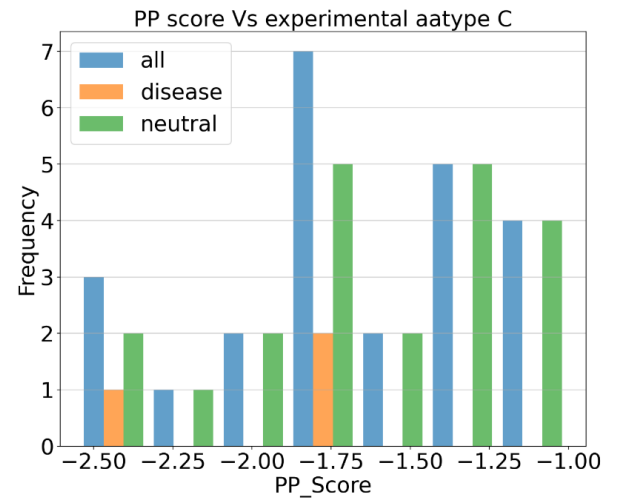
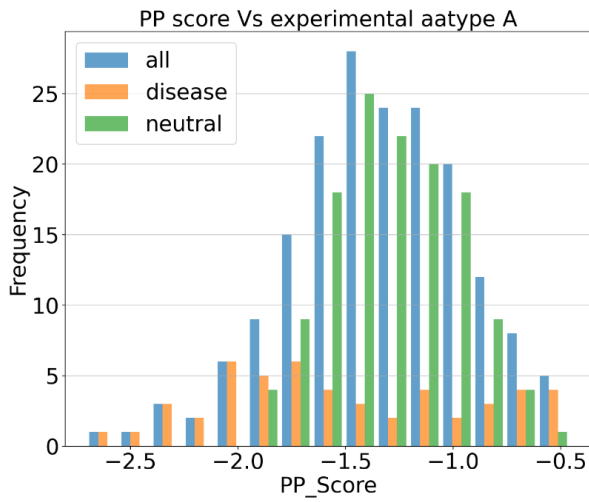
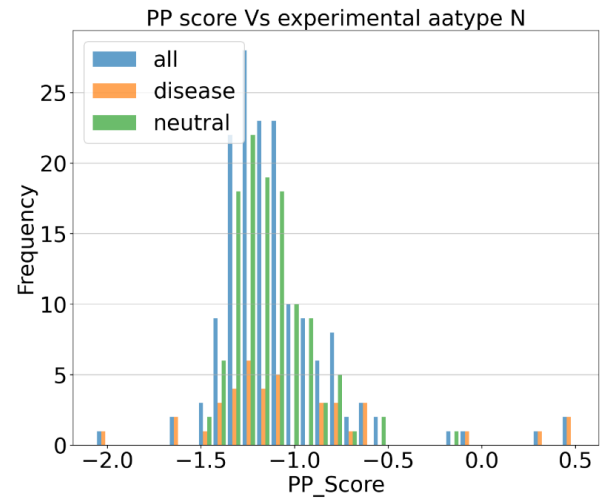
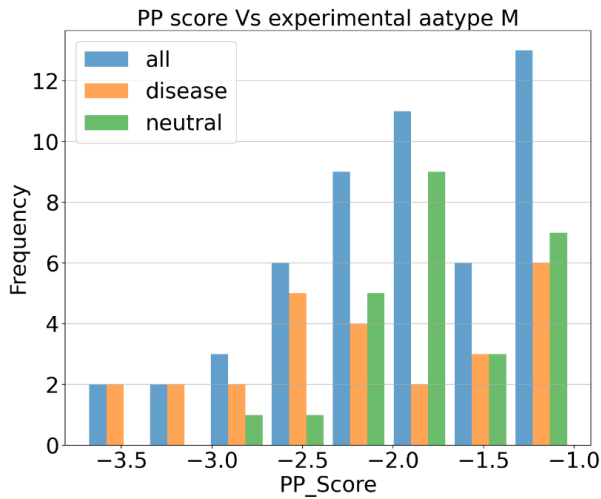
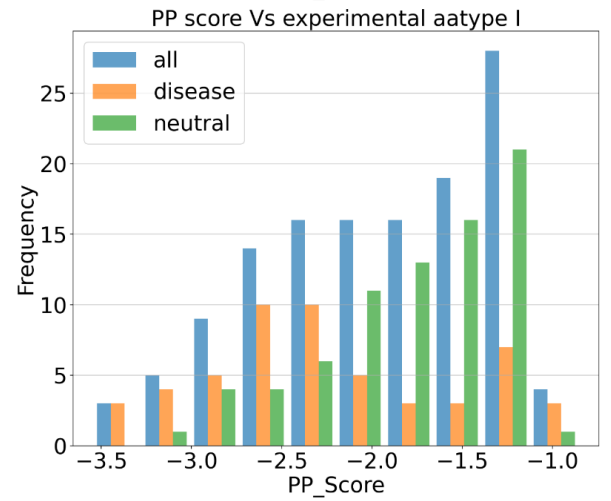
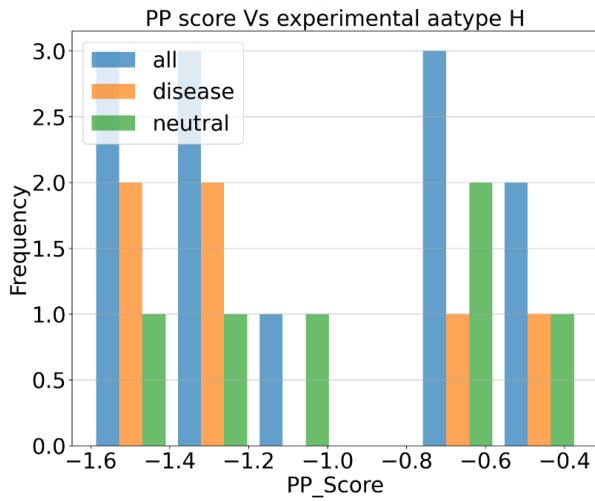
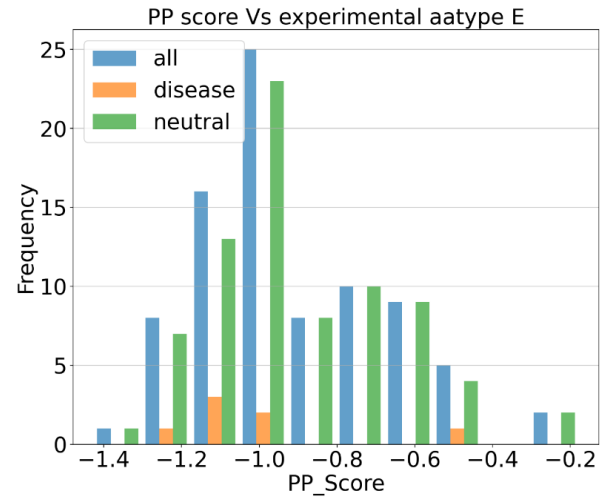
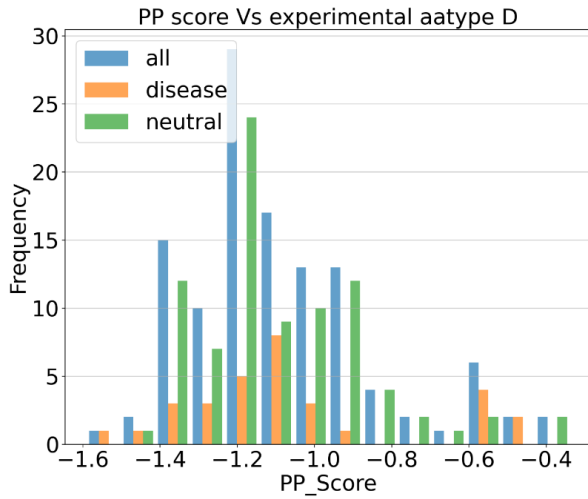


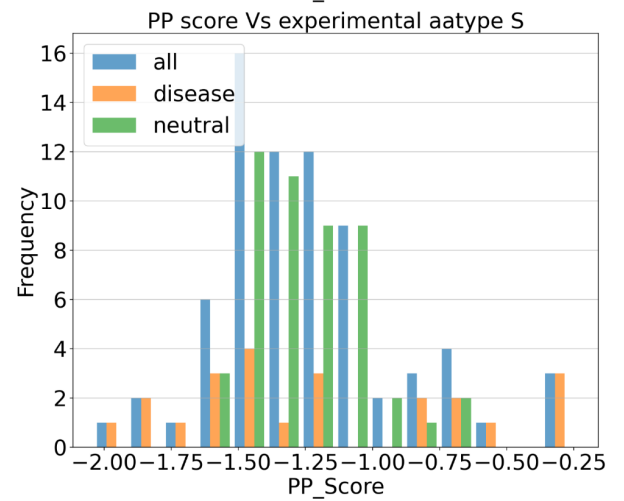
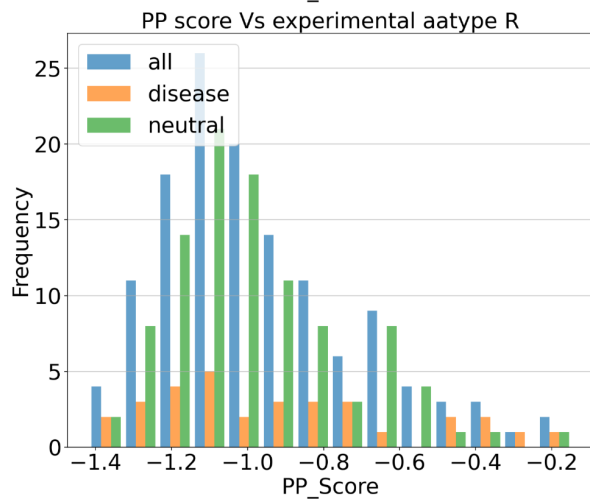
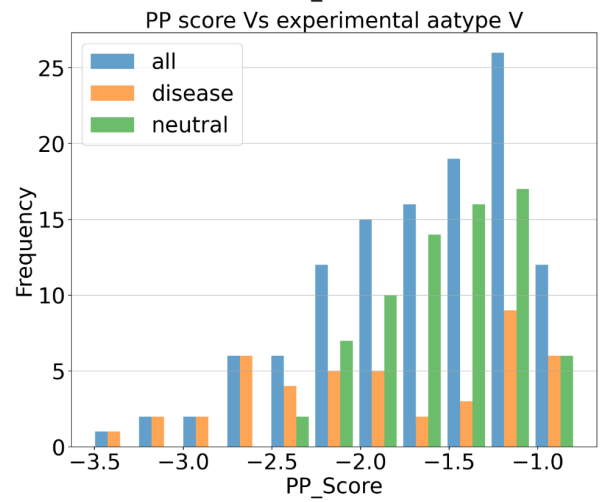
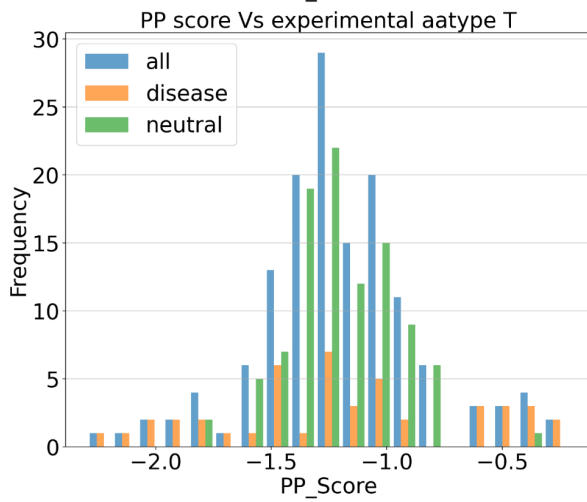
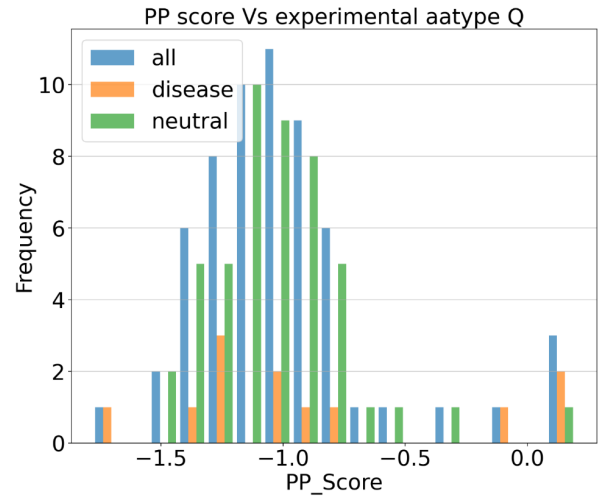
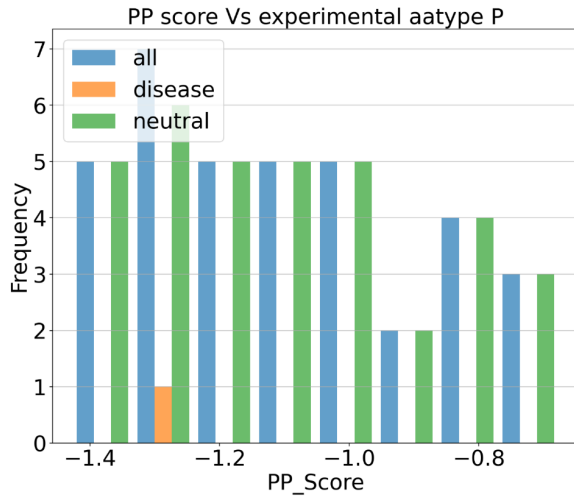
Figure 6: Histogram of FADHM score categorized according to the type of mutated amino acid. Deleterious are coloured in orange, neutral in green and deleterious combined with neutral is colored in blue. Blank plots indicate no wildtype residue was mutated to this particular amino acid.

- Next we checked if there is any such trend when we look at the data based on the wildtype amino acid (ignoring what it is mutated to). We again checked this on the three scores independently (Figure 7, 8 and 9). For the statistical potential, we observed that amino acids A, F, G, L, M, N, S, T, V and W had only deleterious mutations on the left tail of the distribution. Residues like C, and E had all neutral mutations towards the right side of the distribution (Figure 7). For Shannon entropy score, residues like E, F and I showed only neutral mutations between scores of 3 to 4 (Figure 8). For FADHM, residues like C, F and W score a positive score when the mutations are neutral

(Figure 9).







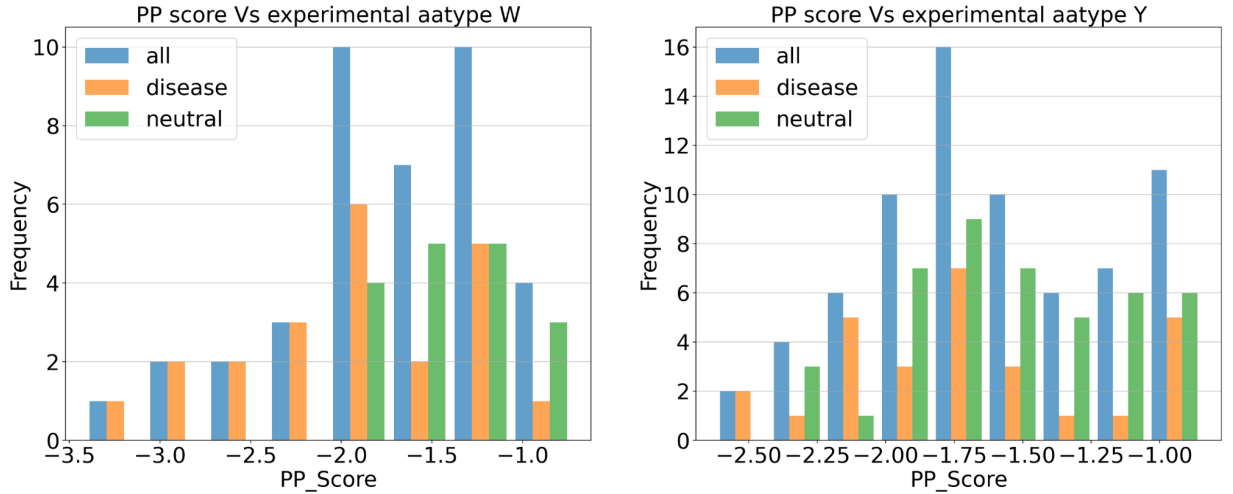
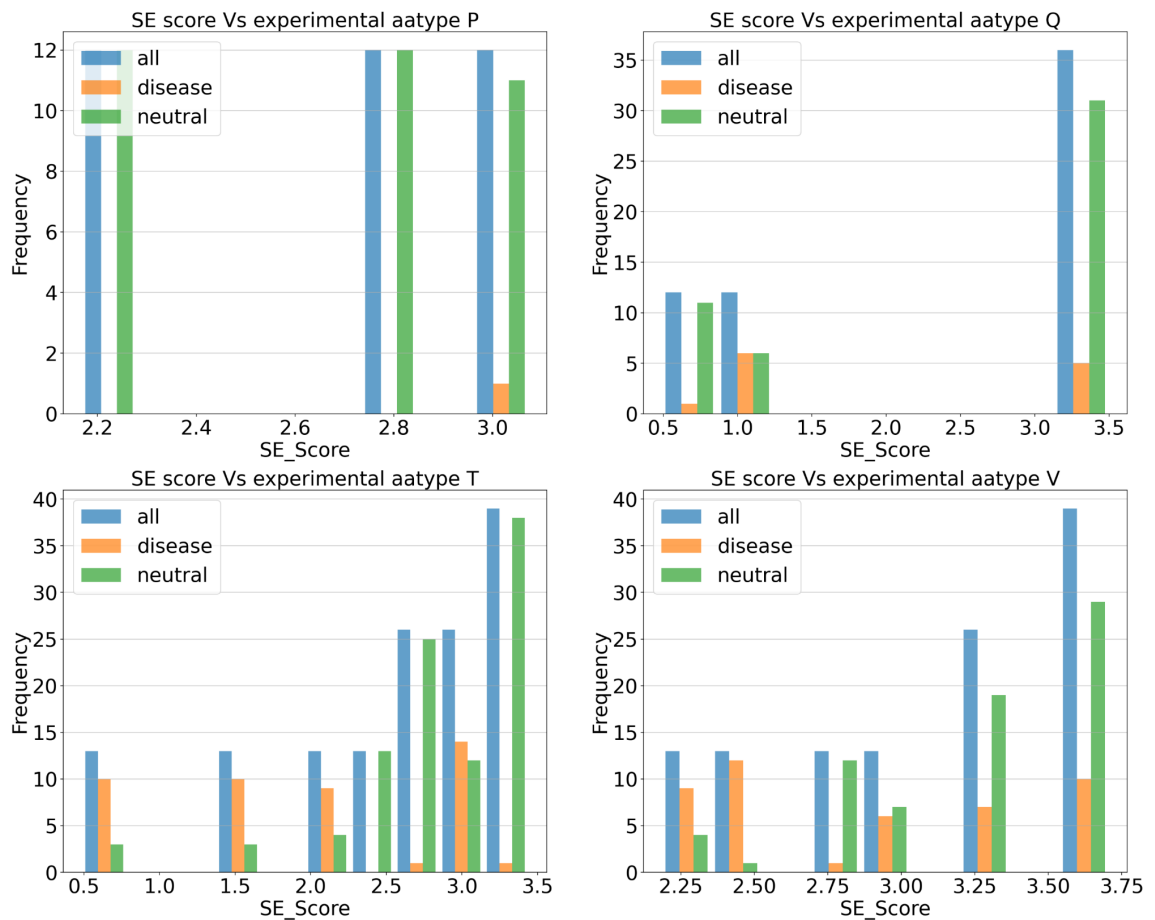
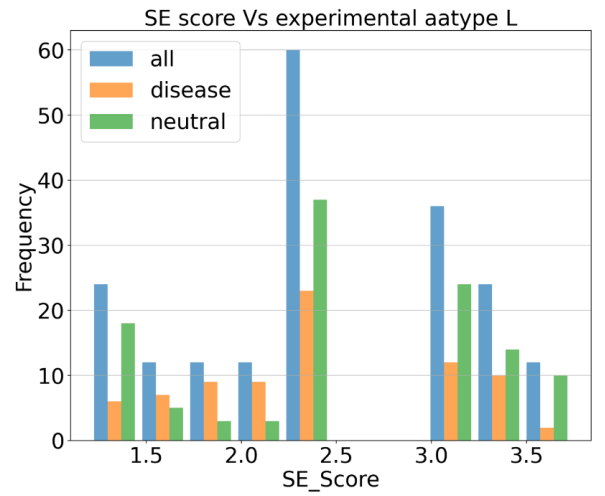
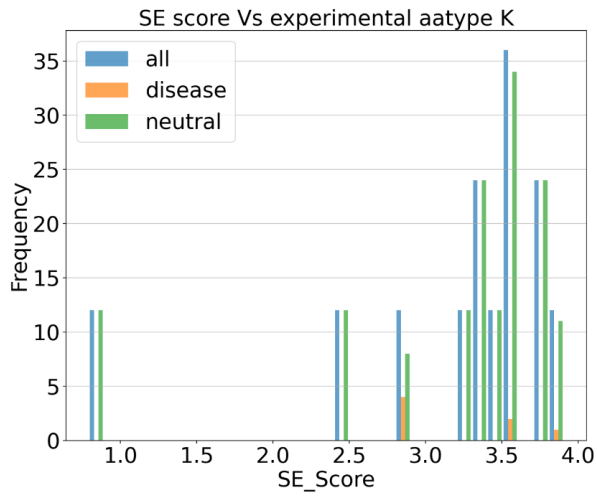
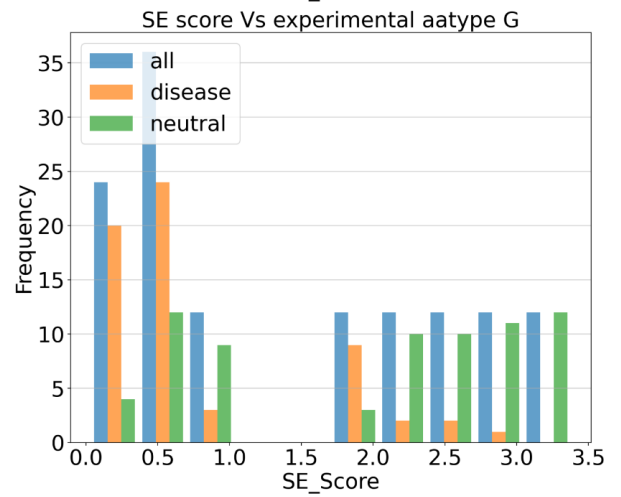
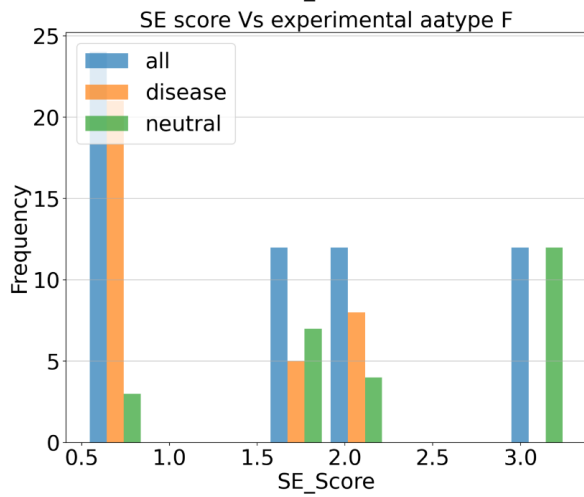
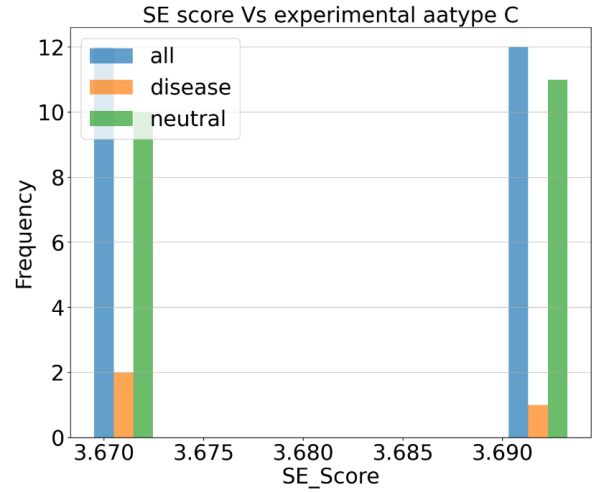
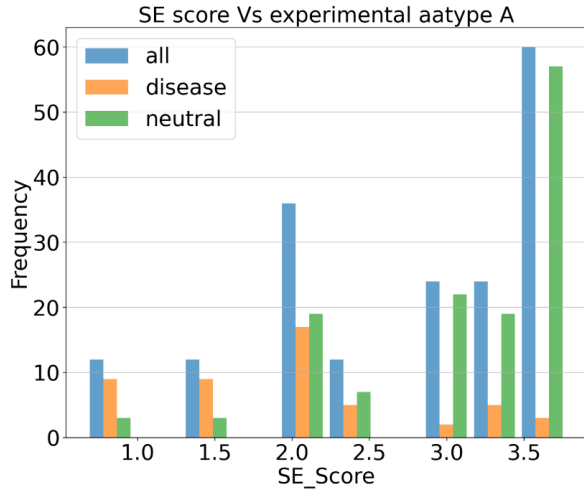
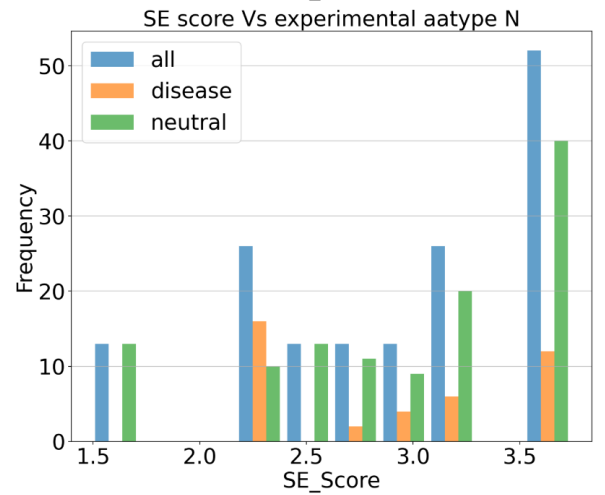
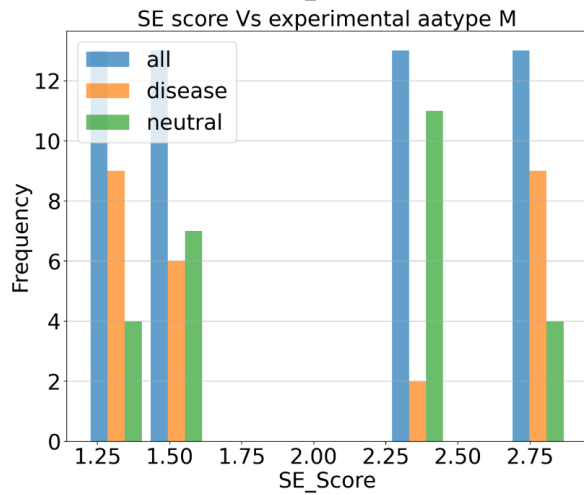
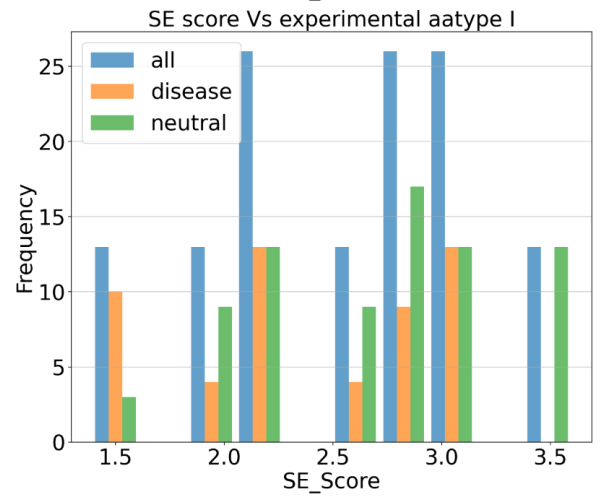
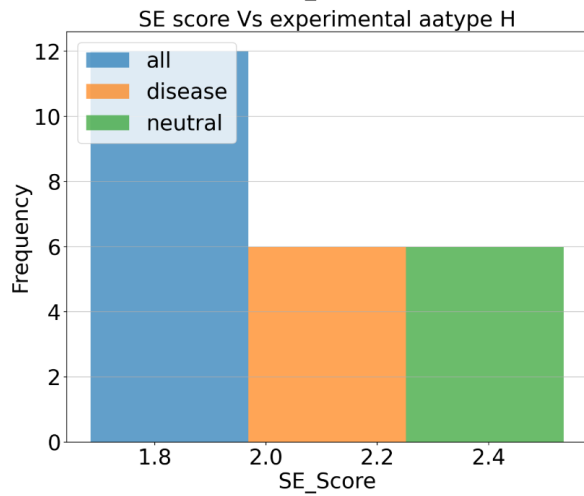
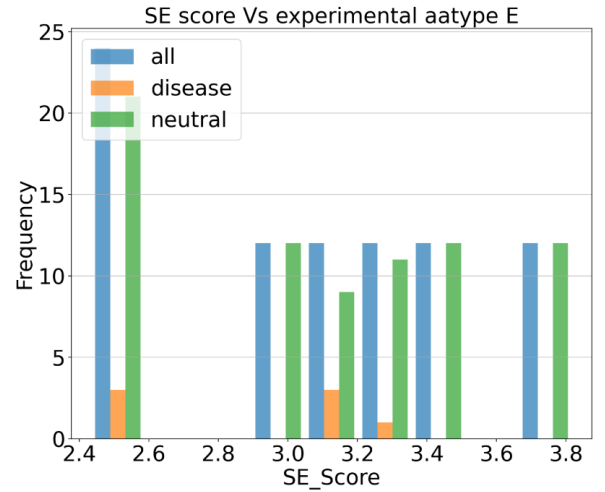
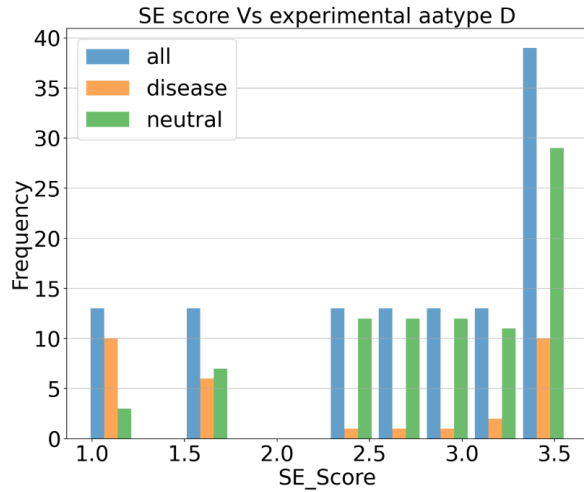


Figure 7: Histogram of statistical potential score categorized according to the type of wildtype amino acid. Deleterious are coloured in orange, neutral in green and deleterious combined with neutral is colored in blue.









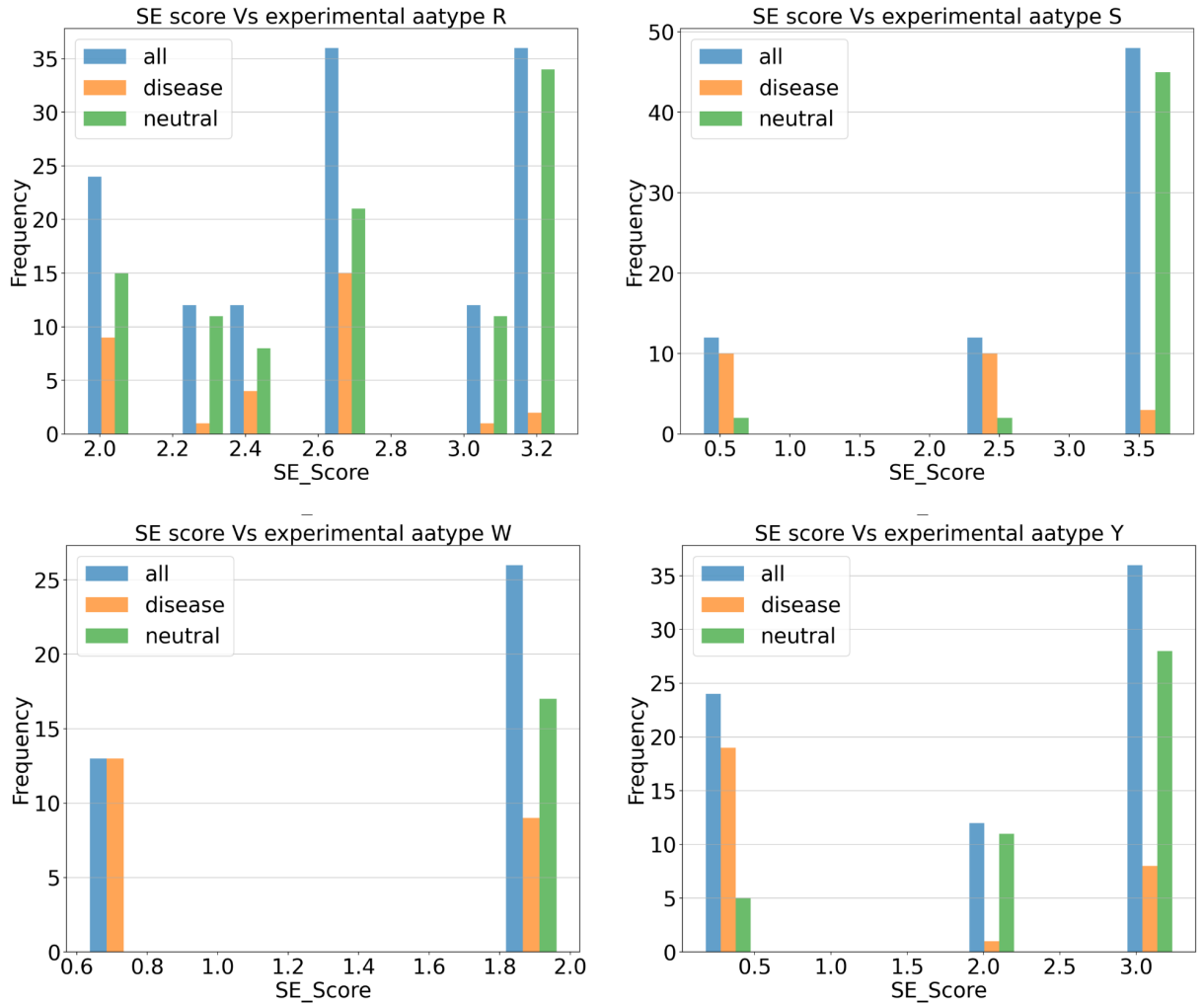
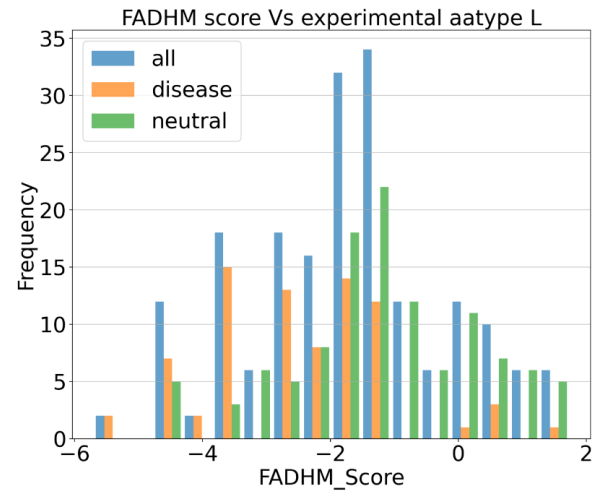
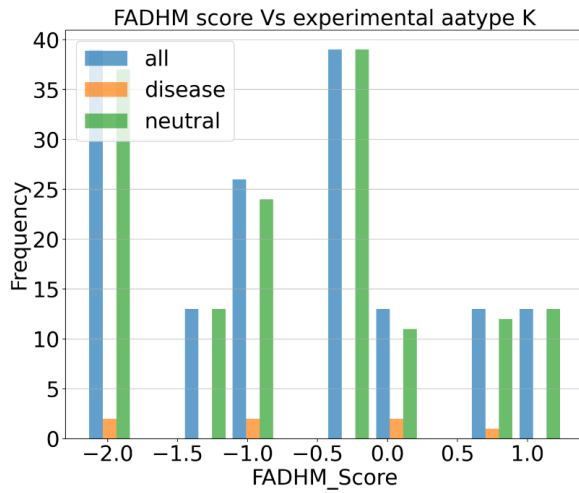
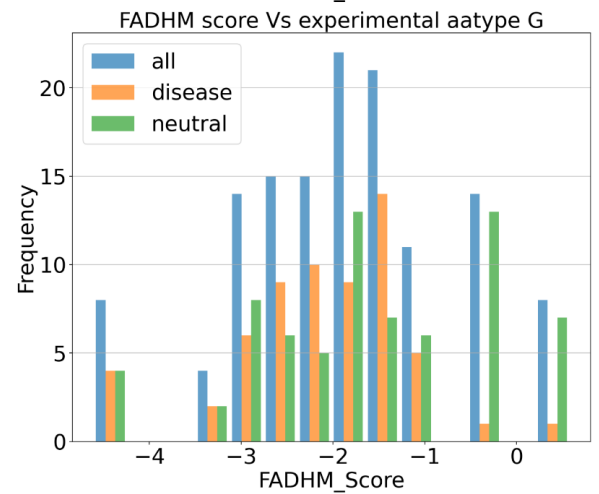
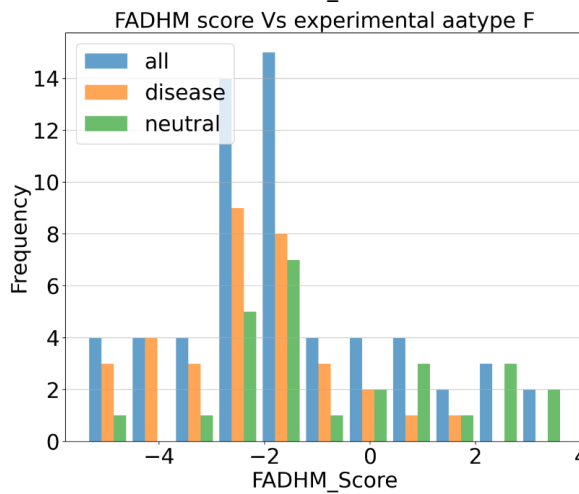
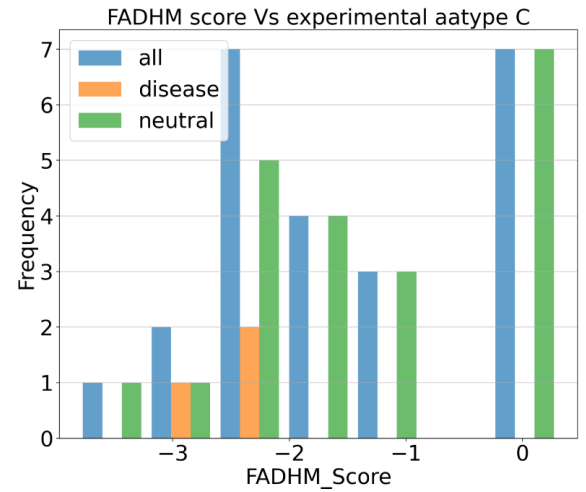
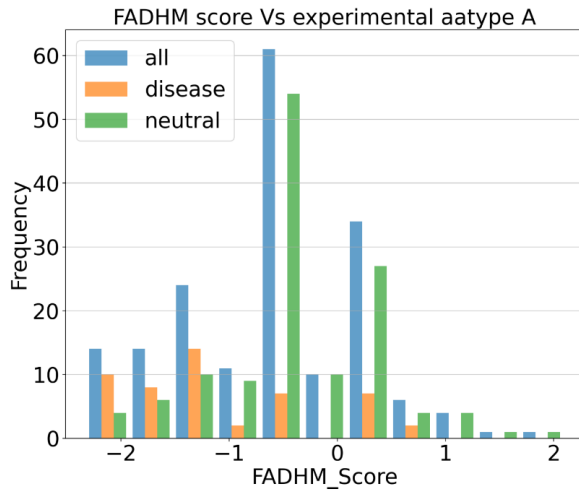
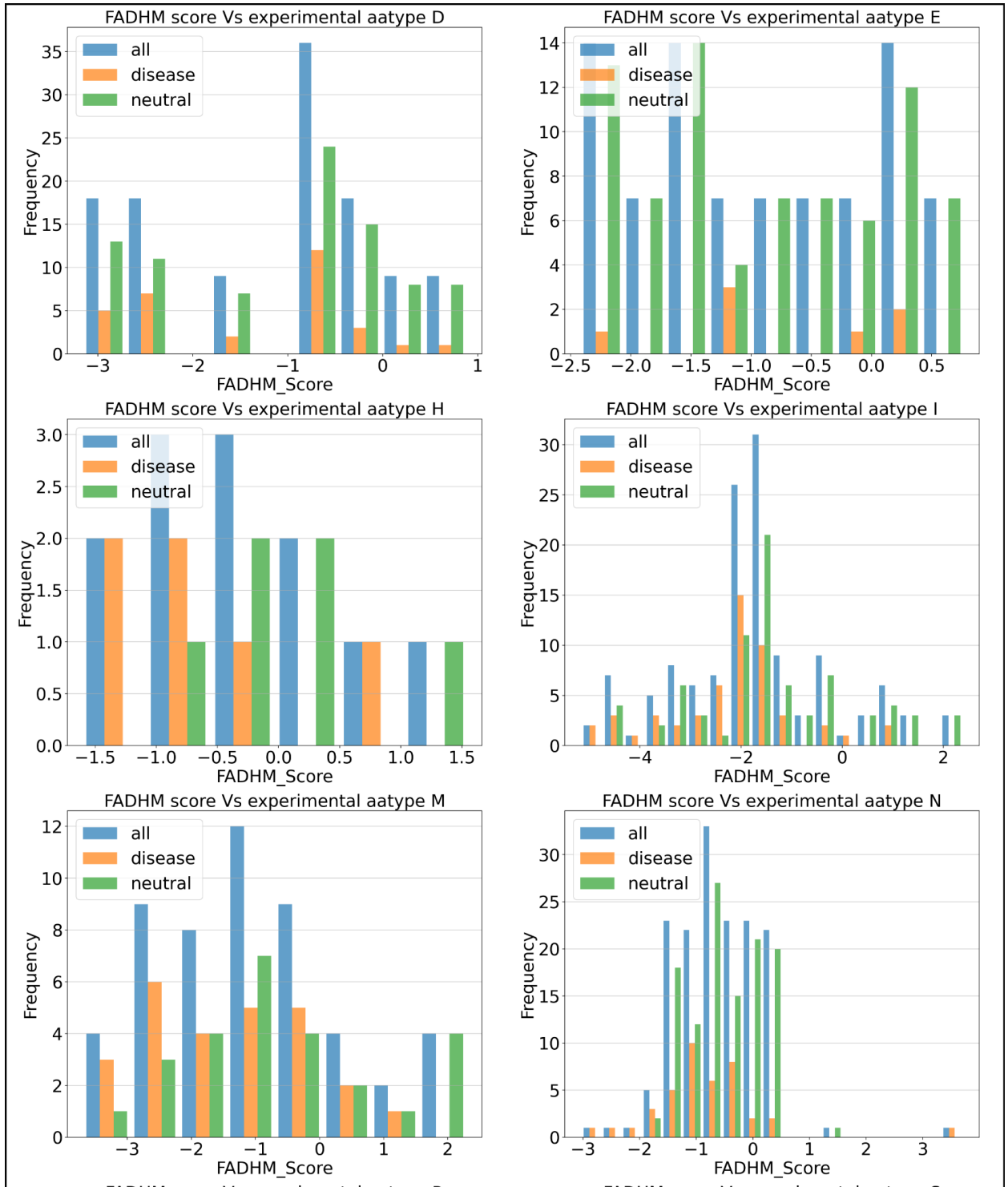
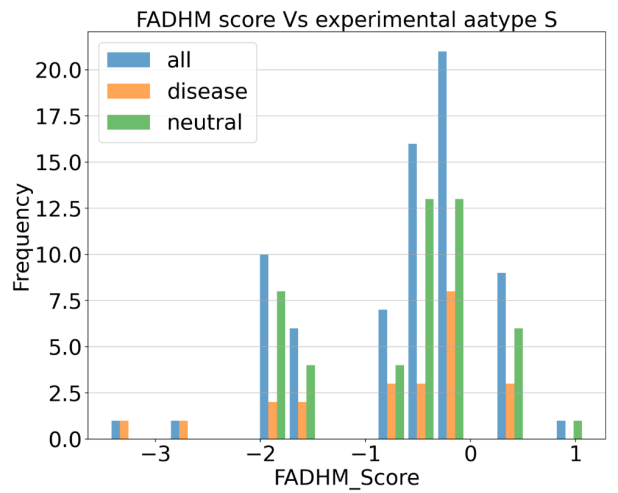
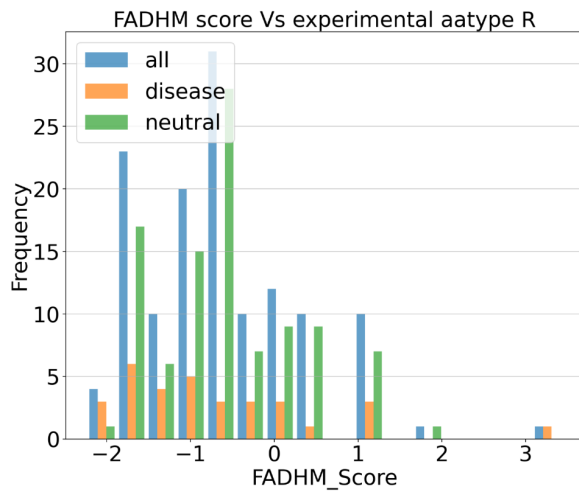
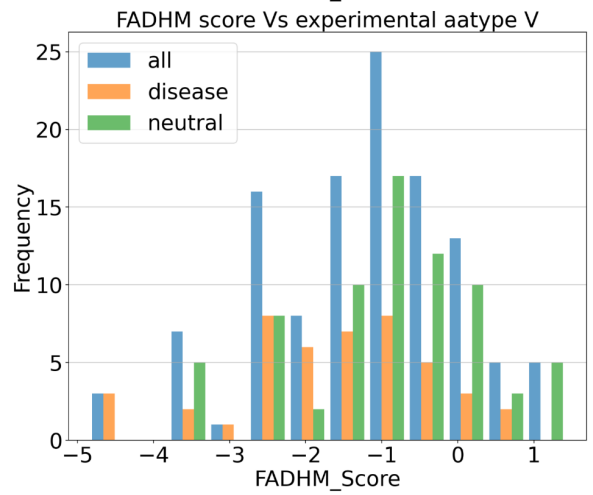
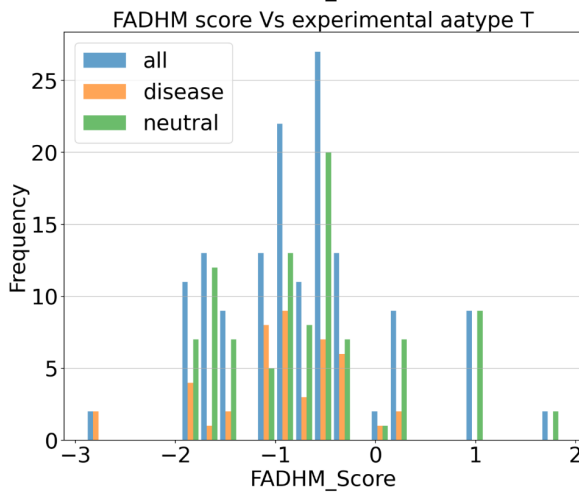
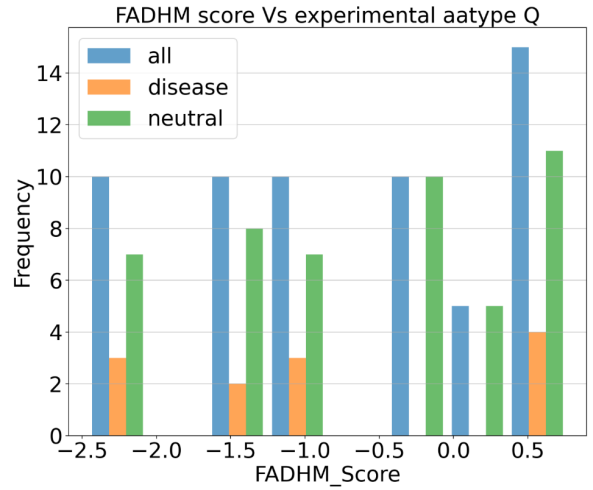
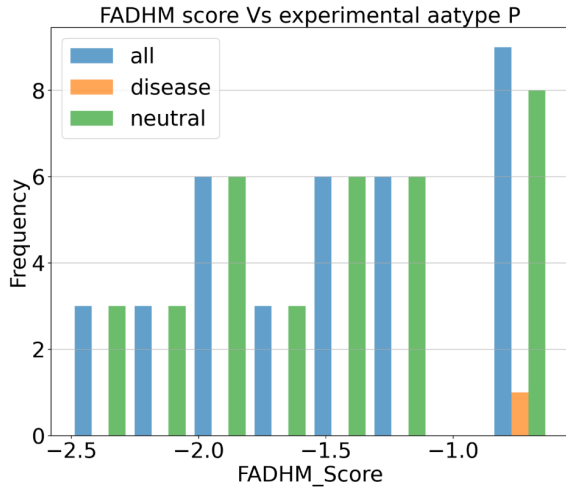


Figure 8: Histogram of SE score categorized according to the type of wildtype amino acid. Deleterious are coloured in orange, neutral in green and deleterious combined with neutral is colored in blue.







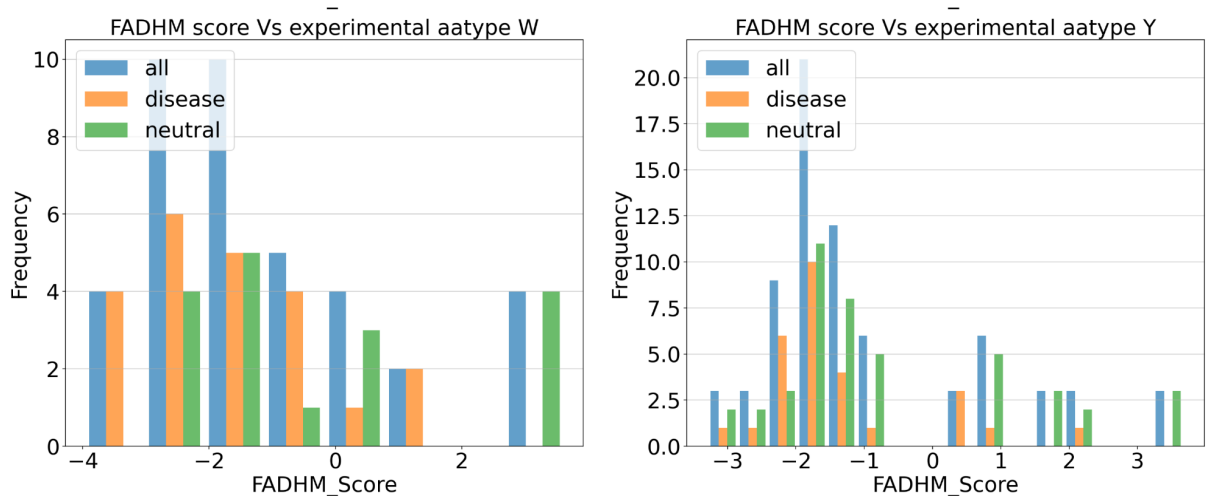


Figure 9: Histogram of FADHM score categorized according to the type of wildtype amino acid. Deleterious are coloured in orange, neutral in green and deleterious combined with neutral is colored in blue.

Since we observed trends when we categorized data based on various properties, we decided to combine the three scores, along with associated properties like depth, amino acid mutation details, etc (Refer to chapter 4 for more details).

## Publications

1. Sen N<sup>1</sup>, Kanitkar TR<sup>1</sup>, Roy AA<sup>1</sup>, Soni N, Amritkar K, Supekar S, Nair S, Singh G, Madhusudhan MS. Predicting and designing therapeutics against the Nipah virus. PLoS Negl Trop Dis. 2019 Dec 12;13(12), e0007419. doi: 10.1371/journal.pntd.0007419. PMID: 31830030; PMCID: PMC6907750.
2. Kanitkar TR, Sen N, Nair S, Soni N, Amritkar K, Ramtirtha Y, Madhusudhan MS. Methods for Molecular Modelling of Protein Complexes. Methods Mol Biol. 2021;2305:53-80. doi: 10.1007/978-1-0716-1406-8@3. PMID: 33950384.
3. Tan KP<sup>1</sup>, Kanitkar TR<sup>1</sup>, Kwoh CK, M.S.Madhusudhan. Packpred, Predicting the Functional Effect of Missense Mutations. Front Mol Biosci. 2021 Aug 20;8:646288. doi: 10.3389/fmolb.2021.646288. PMID: 34490344; PMCID: PMC8417552.
4. Ferro R, Carroll A, Pereira AM, Reen V, Stojiljkovic A, Prince C, Janghra N, Roxanis I, Gazinska P, Annunziato S, Jonkers J, Kanitkar TR, Patel N, Liv N, Alexander J, Quist J, Pardo M, Roumeliotis TI, Choudhary JS, Weekes D, Marra P, Natrajan R, Grigoriadis A, Madhusudhan MS, Haider S, Lord CJ, Tutt A. GPR89 – a novel oncogene regulating unfolded protein response via endoplasmic reticulum pH (**Submitted**)
5. Kanitkar TR and M.S.Madhusudhan. Predicting the molecular mechanism of Class A GPCR activation. (**Under preparation**)
6. Jadhav A, Kanitkar TR and M.S.Madhusudhan. Designing Therapeutics using Peptide Inhibitors against the SARS-CoV-2. (**Under preparation**)

# References

1. Biochemistry, Primary Protein Structure - StatPearls - NCBI Bookshelf. <https://www.ncbi.nlm.nih.gov/books/NBK564343/>. Accessed 16 Feb 2023
2. Muhammed MT, Aki-Yalcin E (2019) Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chem Biol Drug Des* 93:12–20. <https://doi.org/10.1111/cbdd.13388>
3. Nguyen MN, Tan KP, Madhusudhan MS (2011) CLICK - Topology-independent comparison of biomolecular 3D structures. *Nucleic Acids Res* 39:W24–W28. <https://doi.org/10.1093/nar/gkr393>
4. Nguyen MN, Madhusudhan MS (2011) Biological insights from topology independent comparison of protein 3D structures. *Nucleic Acids Res* 39:e94–e94. <https://doi.org/10.1093/nar/gkr348>
5. Shen M, Sali A (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci* 15:2507–2524. <https://doi.org/10.1110/ps.062416606>
6. Structure\_comparison\_and\_alignment\_chapter2\_matri\_renom
7. Diamond R (1976) On the comparison of conformations using linear and quadratic transformations. *Acta Crystallogr Sect A* 32:1–10. <https://doi.org/10.1107/S0567739476000016>
8. Kearsley SK (1989) On the orthogonal transformation used for structural comparisons. *Acta Crystallogr Sect A* 45:208–210. <https://doi.org/10.1107/S0108767388010128>
9. Kabsch W (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr Sect A* 32:922–923. <https://doi.org/10.1107/S0567739476001873>
10. Hendlich M, Rippmann F, Barnickel G (1997) LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model* 15:359–363. [https://doi.org/10.1016/S1093-3263\(98\)00002-3](https://doi.org/10.1016/S1093-3263(98)00002-3)
11. Laskowski RA (1995) SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* 13:323–330. [https://doi.org/10.1016/0263-7855\(95\)00073-9](https://doi.org/10.1016/0263-7855(95)00073-9)
12. Armon A, Graur D, Ben-Tal N (2001) ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol* 307:447–463. <https://doi.org/10.1006/jmbi.2000.4474>
13. Brylinski M, Skolnick J (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci U S A* 105:129–134. <https://doi.org/10.1073/pnas.0707684105>

14. McGreig JE, Uri H, Antczak M, et al (2022) 3DLigandSite: structure-based prediction of protein-ligand binding sites. *Nucleic Acids Res* 50:W13–W20. <https://doi.org/10.1093/nar/gkac250>
15. Meng X-Y, Zhang H-X, Mezei M, Cui M (2012) Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Curr Comput Aided-Drug Des* 7:146–157. <https://doi.org/10.2174/157340911795677602>
16. Kanitkar TR, Sen N, Nair S, et al (2021) Methods for Molecular Modelling of Protein Complexes. In: *Methods in Molecular Biology*. Humana Press Inc., pp 53–80
17. Sliwoski G, Kothiwale S, Meiler J, Lowe EW (2014) Computational methods in drug discovery. *Pharmacol. Rev.* 66:334–395
18. Morris GM, Huey R, Lindstrom W, et al (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* 30:2785–2791. <https://doi.org/10.1002/jcc.21256>
19. Allen WJ, Balias TE, Mukherjee S, et al (2015) DOCK 6: Impact of new features and current docking performance. *J Comput Chem* 36:1132–1156. <https://doi.org/10.1002/jcc.23905>
20. Grosdidier A, Zoete V, Michielin O (2011) SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res* 39:. <https://doi.org/10.1093/nar/gkr366>
21. Maia EHB, Assis LC, de Oliveira TA, et al (2020) Structure-Based Virtual Screening: From Classical to Artificial Intelligence. *Front. Chem.* 8
22. Taylor RD, Jewsbury PJ, Essex JW (2002) A review of protein-small molecule docking methods. *J Comput Aided Mol Des* 16:151–166. <https://doi.org/10.1023/A:1020155510718>
23. Pagadala NS, Syed K, Tuszynski J (2017) Software for molecular docking: a review. *Biophys. Rev.* 9:91–102
24. Hollingsworth SA, Dror RO (2018) Molecular Dynamics Simulation for All. *Neuron* 99:1129–1143
25. Lin FY, MacKerell AD (2019) Force Fields for Small Molecules. In: *Methods in Molecular Biology*. Humana Press Inc., pp 21–54
26. Abraham MJ, Murtola T, Schulz R, et al (2015) Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1–2:19–25. <https://doi.org/10.1016/j.softx.2015.06.001>
27. Phillips JC, Hardy DJ, Maia JDC, et al (2020) Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J Chem Phys* 153:. <https://doi.org/10.1063/5.0014475>
28. Pearlman DA, Case DA, Caldwell JW, et al (1995) AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy



- calculations to simulate the structural and energetic properties of molecules. *Comput Phys Commun* 91:1–41. [https://doi.org/10.1016/0010-4655\(95\)00041-D](https://doi.org/10.1016/0010-4655(95)00041-D)
29. Durrant JD, McCammon JA (2011) Molecular dynamics simulations and drug discovery. *BMC Biol.* 9:71
  30. Sippl MJ (1995) Knowledge-based potentials for proteins. *Curr Opin Struct Biol* 5:229–235. [https://doi.org/10.1016/0959-440X\(95\)80081-6](https://doi.org/10.1016/0959-440X(95)80081-6)
  31. Poole AM, Ranganathan R (2006) Knowledge-based potentials in protein design. *Curr. Opin. Struct. Biol.* 16:508–513
  32. Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins Struct Funct Bioinforma* 17:355–362. <https://doi.org/10.1002/prot.340170404>
  33. Zhou H, Skolnick J (2011) GOAP: A generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys J* 101:2043–2052. <https://doi.org/10.1016/j.bpj.2011.09.012>
  34. Park J, Saitou K (2014) ROTAS: A rotamer-dependent, atomic statistical potential for assessment and prediction of protein structures. *BMC Bioinformatics* 15:1–16. <https://doi.org/10.1186/1471-2105-15-307>
  35. Chao J, Tang F, Xu L (2022) Developments in Algorithms for Sequence Alignment: A Review. *Biomolecules* 12
  36. Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
  37. Altschul SF, Madden TL, Schäffer AA, et al (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402
  38. Shannon CE (1948) A Mathematical Theory of Communication. *Bell Syst Tech J* 27:379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
  39. Stewart JJ, Lee CY, Ibrahim S, et al (1997) A Shannon entropy analysis of immunoglobulin and T cell receptor. *Mol Immunol* 34:1067–1082. [https://doi.org/10.1016/S0161-5890\(97\)00130-2](https://doi.org/10.1016/S0161-5890(97)00130-2)
  40. Lee B, Richards FM (1971) The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol* 55:. [https://doi.org/10.1016/0022-2836\(71\)90324-X](https://doi.org/10.1016/0022-2836(71)90324-X)
  41. Chakravarty S, Varadarajan R (1999) Residue depth: A novel parameter for the analysis of protein structure and stability. *Structure* 7:723–732. [https://doi.org/10.1016/S0969-2126\(99\)80097-5](https://doi.org/10.1016/S0969-2126(99)80097-5)
  42. Tan KP, Varadarajan R, Madhusudhan MS (2011) DEPTH: A web server to compute depth and predict small-molecule binding cavities in proteins. *Nucleic Acids Res* 39:W242–W248. <https://doi.org/10.1093/nar/gkr356>
  43. Latorraca NR, Venkatakrisnan AJ, Dror RO (2017) GPCR dynamics: Structures in motion. *Chem*

Rev 117:139–155. <https://doi.org/10.1021/acs.chemrev.6b00177>

44. Wright PT, Schobesberger S, Gorelik J (2015) Studying GPCR/cAMP pharmacology from the perspective of cellular structure. *Front. Pharmacol.* 6
45. Sun L, Ye RD (2012) Role of G protein-coupled receptors in inflammation. *Acta Pharmacol. Sin.* 33:342–350
46. Lämmermann T, Kastenmüller W (2019) Concepts of GPCR-controlled navigation in the immune system. *Immunol. Rev.* 289:205–231
47. Julius D, Nathans J (2012) Signaling by sensory receptors. *Cold Spring Harb Perspect Biol* 4:. <https://doi.org/10.1101/cshperspect.a005991>
48. Dalesio NM, Barreto Ortiz SF, Pluznick JL, Berkowitz DE (2018) Olfactory, taste, and photo sensory receptors in non-sensory organs: It just makes sense. *Front Physiol* 9:1–19. <https://doi.org/10.3389/fphys.2018.01673>
49. Oliveira de Souza C, Sun X, Oh D (2021) Metabolic Functions of G Protein-Coupled Receptors and  $\beta$ -Arrestin-Mediated Signaling Pathways in the Pathophysiology of Type 2 Diabetes and Obesity. *Front. Endocrinol. (Lausanne)*. 12
50. Yang D, Zhou Q, Labroska V, et al (2021) G protein-coupled receptors: structure- and function-based drug discovery. *Signal Transduct Target Ther* 6:. <https://doi.org/10.1038/s41392-020-00435-w>
51. Rasmussen SGF, Devree BT, Zou Y, et al (2011) Crystal structure of the  $\beta$  2 adrenergic receptor-Gs protein complex. *Nature* 477:549–557. <https://doi.org/10.1038/nature10361>
52. Filipek S (2019) Molecular switches in GPCRs. *Curr Opin Struct Biol* 55:114–120. <https://doi.org/10.1016/j.sbi.2019.03.017>
53. Katritch V, Fenalti G, Abola EE, et al (2014) Allosteric sodium in class A GPCR signaling. *Trends Biochem. Sci.* 39:233–244
54. Fritze O, Filipek S, Kuksa V, et al (2003) Role of the conserved NPxxY(x)5,6F motif in the rhodopsin ground state and during activation. *Proc Natl Acad Sci U S A* 100:2290–2295. <https://doi.org/10.1073/pnas.0435715100>
55. Schönege AM, Gallion J, Picard LP, et al (2017) Evolutionary action and structural basis of the allosteric switch controlling  $\beta$ 2AR functional selectivity. *Nat Commun* 8:1–12. <https://doi.org/10.1038/s41467-017-02257-x>
56. Favre N, Fanelli F, Missotten M, et al (2005) The DRY motif as a molecular switch of the human oxytocin receptor. *Biochemistry* 44:9990–10008. <https://doi.org/10.1021/bi0509853>
57. Audet M, Bouvier M (2012) Restructuring G-Protein- Coupled Receptor Activation. *Cell* 151:14–23

58. Hilger D (2021) The role of structural dynamics in GPCR-mediated signaling. *FEBS J* 288:2461–2489. <https://doi.org/10.1111/febs.15841>
59. Pin JP, Galvez T, Prézeau L (2003) Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors. *Pharmacol. Ther.* 98:325–354
60. Basith S, Cui M, Macalino SJY, et al (2018) Exploring G protein-coupled receptors (GPCRs) ligand space via cheminformatics approaches: Impact on rational drug design. *Front. Pharmacol.* 9
61. Bortolato A, Doré AS, Hollenstein K, et al (2014) Structure of Class B GPCRs: New horizons for drug discovery. *Br. J. Pharmacol.* 171:3132–3145
62. Chun L, Zhang WH, Liu JF (2012) Structure and ligand recognition of class C GPCRs. *Acta Pharmacol. Sin.* 33:312–323
63. Velazhahan V, Ma N, Pándy-Szekeres G, et al (2021) Structure of the class D GPCR Ste2 dimer coupled to two G proteins. *Nature* 589:148–153. <https://doi.org/10.1038/s41586-020-2994-1>
64. Weis WI, Kobilka BK (2018) The Molecular Basis of G Protein-Coupled Receptor Activation. *Annu. Rev. Biochem.* 87:897–919
65. Okada T, Sugihara M, Bondar AN, et al (2004) The retinal conformation and its environment in rhodopsin in light of a new 2.2 Å crystal structure. *J Mol Biol* 342:571–583. <https://doi.org/10.1016/j.jmb.2004.07.044>
66. Blankenship E, Vahedi-Faridi A, Lodowski DT (2015) The High-Resolution Structure of Activated Opsin Reveals a Conserved Solvent Network in the Transmembrane Region Essential for Activation. *Structure* 23:2358–2364. <https://doi.org/10.1016/j.str.2015.09.015>
67. Cherezov V, Rosenbaum DM, Hanson MA, et al (2007) High-resolution crystal structure of an engineered human  $\beta$ 2-adrenergic G protein-coupled receptor. *Science* (80- ) 318:1258–1265. <https://doi.org/10.1126/science.1150577>
68. Ring AM, Manglik A, Kruse AC, et al (2013) Adrenaline-activated structure of  $\beta$ 2-adrenoceptor stabilized by an engineered nanobody. *Nature* 502:575–579. <https://doi.org/10.1038/nature12572>
69. Suno R, Lee S, Maeda S, et al (2018) Structural insights into the subtype-selective antagonist binding to the M2 muscarinic receptor. *Nat Chem Biol* 14:1150–1158. <https://doi.org/10.1038/s41589-018-0152-y>
70. Kruse AC, Ring AM, Manglik A, et al (2013) Activation and allosteric modulation of a muscarinic acetylcholine receptor. *Nature* 504:101–106. <https://doi.org/10.1038/nature12735>
71. Huang W, Manglik A, Venkatakrisnan AJ, et al (2015) Structural insights into  $\mu$ -opioid receptor activation. *Nature* 524:315–321. <https://doi.org/10.1038/nature14886>
72. Manglik A, Kruse AC, Kobilka TS, et al (2012) Crystal structure of the  $\mu$ -opioid receptor bound to a morphinan antagonist. *Nature* 485:321–326. <https://doi.org/10.1038/nature10954>

73. Weinert T, Olieric N, Cheng R, et al (2017) Serial millisecond crystallography for routine room-temperature structure determination at synchrotrons. *Nat Commun* 8:1–11. <https://doi.org/10.1038/s41467-017-00630-4>
74. White KL, Eddy MT, Gao ZG, et al (2018) Structural Connection between Activation Microswitch and Allosteric Sodium Site in GPCR Signaling. *Structure* 26:259–269.e5. <https://doi.org/10.1016/j.str.2017.12.013>
75. Wu H, Wacker D, Mileni M, et al (2012) Structure of the human  $\kappa$ -opioid receptor in complex with JDTic. *Nature* 485:327–332. <https://doi.org/10.1038/nature10939>
76. Munk C, Isberg V, Mordalski S, et al (2016) GPCRdb: the G protein-coupled receptor database – an introduction. *Br J Pharmacol* 16:2195–2207. <https://doi.org/10.1111/bph.13509>
77. Eswar N, Webb B, Marti-Renom MA, et al (2006) Comparative Protein Structure Modeling Using Modeller. *Curr Protoc Bioinforma* 15:Unit. <https://doi.org/10.1002/0471250953.bi0506s15>
78. Šali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234:779–815. <https://doi.org/10.1006/jmbi.1993.1626>
79. Zhang Y, Skolnick J (2005) TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 33:2302–2309. <https://doi.org/10.1093/nar/gki524>
80. Pronk S, Páll S, Schulz R, et al (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29:845–854. <https://doi.org/10.1093/bioinformatics/btt055>
81. Huang J, Mackerell AD (2013) CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J Comput Chem* 34:2135–2145. <https://doi.org/10.1002/jcc.23354>
82. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: An  $N \cdot \log(N)$  method for Ewald sums in large systems. *J Chem Phys* 98:10089–10092. <https://doi.org/10.1063/1.464397>
83. Berendsen HJC, Postma JPM, Van Gunsteren WF, et al (1984) Molecular dynamics with coupling to an external bath. *J Chem Phys* 81:3684–3690. <https://doi.org/10.1063/1.448118>
84. Parrinello M, Rahman A (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys* 52:7182–7190. <https://doi.org/10.1063/1.328693>
85. Ballesteros JA, Weinstein H (1995) Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods Neurosci* 25:366–428. [https://doi.org/10.1016/S1043-9471\(05\)80049-7](https://doi.org/10.1016/S1043-9471(05)80049-7)
86. Anantkrishnan S, Naganathan AN (2023) Thermodynamic architecture and conformational plasticity of GPCRs. *Nat Commun* 14:1–14. <https://doi.org/10.1038/s41467-023-35790-z>
87. Zhou Q, Yang D, Wu M, et al (2019) Common activation mechanism of class a GPCRs. *Elife* 8:.

<https://doi.org/10.7554/eLife.50279>

88. Rodríguez-Espigares I, Torrens-Fontanals M, Tiemann JKS, et al (2020) GPCRmd uncovers the dynamics of the 3D-GPCRome. *Nat Methods* 17:777–787. <https://doi.org/10.1038/s41592-020-0884-y>
89. Venkatakrisnan AJ, Ma AK, Fonseca R, et al (2019) Diverse GPCRs exhibit conserved water networks for stabilization and activation. *Proc Natl Acad Sci U S A* 116:3288–3293. <https://doi.org/10.1073/pnas.1809251116>
90. Jumper J, Evans R, Pritzel A, et al (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–589. <https://doi.org/10.1038/s41586-021-03819-2>
91. Baek M, DiMaio F, Anishchenko I, et al (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science* (80- ) 373:871–876. <https://doi.org/10.1126/science.abj8754>
92. Zhang Z, Miteva MA, Wang L, Alexov E (2012) Analyzing effects of naturally occurring missense mutations. *Comput. Math. Methods Med.* 2012:15
93. Roach JC, Glusman G, Smit AFA, et al (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* (80- ) 328:636–639. <https://doi.org/10.1126/science.1186802>
94. Stranger BE, Stahl EA, Raj T (2011) Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* 187:367–383
95. Craig Venter J, Adams MD, Myers EW, et al (2001) The sequence of the human genome. *Science* (80- ) 291:1304–1351. <https://doi.org/10.1126/science.1058040>
96. Frazer KA, Ballinger DG, Cox DR, et al (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–861. <https://doi.org/10.1038/nature06258>
97. Altshuler DM, Durbin RM, Abecasis GR, et al (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65. <https://doi.org/10.1038/nature11632>
98. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31:3812–3814. <https://doi.org/10.1093/nar/gkg509>
99. Smith HO, Annau TM, Chandrasegaran S (1990) Finding sequence motifs in groups of functionally related proteins. *Proc Natl Acad Sci U S A* 87:826–830. <https://doi.org/10.1073/pnas.87.2.826>
100. Capriotti E, Fariselli P, Casadio R (2005) I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res* 33:. <https://doi.org/10.1093/nar/gki375>
101. Masso M, Vaisman II (2014) AUTO-MUTE 2.0: A portable framework with enhanced capabilities for predicting protein functional consequences upon mutation. *Adv Bioinformatics* 2014:.

- <https://doi.org/10.1155/2014/278385>
102. Dehouck Y, Grosfils A, Folch B, et al (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics* 25:2537–2543. <https://doi.org/10.1093/bioinformatics/btp445>
  103. Pires DEV, Ascher DB, Blundell TL (2014) MCSM: Predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* 30:335–342. <https://doi.org/10.1093/bioinformatics/btt691>
  104. Worth CL, Preissner R, Blundell TL (2011) SDM - A server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res* 39:W215. <https://doi.org/10.1093/nar/gkr363>
  105. Pandurangan AP, Ochoa-Montaña B, Ascher DB, Blundell TL (2017) SDM: A server for predicting effects of mutations on protein stability. *Nucleic Acids Res* 45:W229–W235. <https://doi.org/10.1093/nar/gkx439>
  106. Pires DEV, Ascher DB, Blundell TL (2014) DUET: A server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res* 42:. <https://doi.org/10.1093/nar/gku411>
  107. Ittisoponpisan S, Islam SA, Khanna T, et al (2019) Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? *J Mol Biol* 431:2197–2212. <https://doi.org/10.1016/j.jmb.2019.04.009>
  108. Rodrigues CHM, Pires DEV, Ascher DB (2020) <scp>DynaMut2</scp> : Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Sci* pro.3942. <https://doi.org/10.1002/pro.3942>
  109. Adzhubei IA, Schmidt S, Peshkin L, et al (2010) A method and server for predicting damaging missense mutations. *Nat. Methods* 7:248–249
  110. Farheen N, Sen N, Nair S, et al (2017) Depth dependent amino acid substitution matrices and their use in predicting deleterious mutations. *Prog Biophys Mol Biol* 128:14–23. <https://doi.org/10.1016/j.pbiomolbio.2017.02.004>
  111. Berman HM, Westbrook J, Feng Z, et al (2000) The Protein Data Bank. *Nucleic Acids Res.* 28:235–242
  112. Rennell D, Bouvier SE, Hardy LW, Poteete AR (1991) Systematic mutation of bacteriophage T4 lysozyme. *J Mol Biol* 222:. [https://doi.org/10.1016/0022-2836\(91\)90738-R](https://doi.org/10.1016/0022-2836(91)90738-R)
  113. Adkar B V., Tripathi A, Sahoo A, et al (2012) Protein model discrimination using mutational sensitivity derived from deep sequencing. *Structure* 20:371–381. <https://doi.org/10.1016/j.str.2011.11.021>
  114. Weaver LH, Matthews BW (1987) Structure of bacteriophage T4 lysozyme refined at 1.7 Å

- resolution. *J Mol Biol* 193:189–199. [https://doi.org/10.1016/0022-2836\(87\)90636-X](https://doi.org/10.1016/0022-2836(87)90636-X)
115. Loris R, Dao-Thi MH, Bahassi EM, et al (1999) Crystal structure of CcdB, a topoisomerase poison from *E. coli*. *J Mol Biol* 285:1667–1677. <https://doi.org/10.1006/jmbi.1998.2395>
  116. Bateman A, Martin MJ, O'Donovan C, et al (2017) UniProt: The universal protein knowledgebase. *Nucleic Acids Res* 45:D158–D169. <https://doi.org/10.1093/nar/gkw1099>
  117. Landrum MJ, Lee JM, Riley GR, et al (2014) ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 42:D980. <https://doi.org/10.1093/nar/gkt1113>
  118. Karczewski KJ, Weisburd B, Thomas B, et al (2017) The ExAC browser: Displaying reference data information from over 60 000 exomes. *Nucleic Acids Res* 45:D840–D845. <https://doi.org/10.1093/nar/gkw971>
  119. Tan KP, Nguyen TB, Patel S, et al (2013) Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. *Nucleic Acids Res* 41:. <https://doi.org/10.1093/nar/gkt503>
  120. Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* 213:859–883. [https://doi.org/10.1016/S0022-2836\(05\)80269-4](https://doi.org/10.1016/S0022-2836(05)80269-4)
  121. Pillai VS, Krishna G, Veetil MV (2020) Nipah virus: Past outbreaks and future containment. *Viruses* 12
  122. Xu K, Chan Y-P, Bradel-Tretheway B, et al (2015) Crystal Structure of the Pre-fusion Nipah Virus Fusion Glycoprotein Reveals a Novel Hexamer-of-Trimers Assembly. *PLOS Pathog* 11:e1005322. <https://doi.org/10.1371/journal.ppat.1005322>
  123. Sun W, McCrory TS, Khaw WY, et al (2014) Matrix Proteins of Nipah and Hendra Viruses Interact with Beta Subunits of AP-3 Complexes. *J Virol* 88:13099–13110. <https://doi.org/10.1128/jvi.02103-14>
  124. Yoneda M, Guillaume V, Sato H, et al (2010) The Nonstructural Proteins of Nipah Virus Play a Key Role in Pathogenicity in Experimentally Infected Animals. *PLoS One* 5:e12709. <https://doi.org/10.1371/journal.pone.0012709>
  125. What is Nipah Virus? | Nipah Virus (NiV) | CDC. <https://www.cdc.gov/vhf/nipah/about/index.html>. Accessed 17 Apr 2023
  126. Pinzi L, Rastelli G (2019) Molecular docking: Shifting paradigms in drug discovery. *Int. J. Mol. Sci.* 20
  127. Irwin JJ, Sterling T, Mysinger MM, et al (2012) ZINC: A Free Tool to Discover Chemistry for Biology. <https://doi.org/10.1021/ci3001277>

128. Irwin JJ, Shoichet BK ZINC-A Free Database of Commercially Available Compounds for Virtual Screening
129. Hussein HA, Borrel A, Geneix C, et al (2015) PockDrug-Server: a new web server for predicting pocket druggability on holo and apo proteins. *Nucleic Acids Res* 43:W436--W442. <https://doi.org/10.1093/nar/gkv462>
130. Xu Y, Wang S, Hu Q, et al (2018) CavityPlus: a web server for protein cavity detection with pharmacophore modelling, allosteric site identification and covalent ligand binding ability prediction. *Nucleic Acids Res* 46:W374--W379. <https://doi.org/10.1093/nar/gky380>
131. Pettersen EF, Goddard TD, Huang CC, et al (2004) UCSF Chimera-A visualization system for exploratory research and analysis. *J Comput Chem* 25:1605–1612. <https://doi.org/10.1002/jcc.20084>
132. Wang J, Wolf RM, Caldwell JW, et al (2004) Development and testing of a general Amber force field. *J Comput Chem* 25:1157–1174. <https://doi.org/10.1002/jcc.20035>
133. Wang J, Wang W, Kollman PA, Case DA (2006) Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model* 25:247–260. <https://doi.org/10.1016/j.jmgm.2005.12.005>
134. Maharana J, Patra MC, De BC, et al (2014) Structural insights into the MDP binding and CARD-CARD interaction in zebrafish (*Danio rerio*) NOD2: A molecular dynamics approach. *J Mol Recognit* 27:260–275. <https://doi.org/10.1002/jmr.2357>
135. Dapiaggi F, Pieraccini S, Potenza D, et al (2017) Computer aided design and NMR characterization of an oligopeptide targeting the Ebola virus VP24 protein. *New J Chem* 41:4308–4315. <https://doi.org/10.1039/c6nj04014d>
136. Decherchi S, Berteotti A, Bottegoni G, et al (2015) The ligand binding mechanism to purine nucleoside phosphorylase elucidated via molecular dynamics and machine learning. *Nat Commun* 6. <https://doi.org/10.1038/ncomms7155>
137. Nguyen MN, Sen N, Lin M, et al (2019) Discovering Putative Protein Targets of Small Molecules: A Study of the p53 Activator Nutlin. *J Chem Inf Model* 59:1529–1546. <https://doi.org/10.1021/acs.jcim.8b00762>
138. Zoete V, Cuendet MA, Grosdidier A, Michielin O (2011) SwissParam: A fast force field generation tool for small organic molecules. *J Comput Chem* 32:2359–2368. <https://doi.org/10.1002/jcc.21816>
139. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) LINCS: A linear constraint solver for molecular simulations. *J Comput Chem* 18:1463–1472. [https://doi.org/10.1002/\(SICI\)1096-987X\(199709\)18:12<1463::AID-JCC4>3.0.CO;2-H](https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H)
140. Kollman PA, Massova I, Reyes C, et al (2000) Calculating Structures and Free Energies of



- Complex Molecules: Combining Molecular Mechanics and Continuum Models. <https://doi.org/10.1021/AR000033J>
141. Srinivasan J, Cheatham TE, Cieplak P, et al (1998) Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate–DNA Helices. *J Am Chem Soc* 120:. <https://doi.org/10.1021/JA981844+>
  142. Baker NA, Sept D, Joseph S, et al (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* 98:10037–10041. <https://doi.org/10.1073/pnas.181342398>
  143. Papissoni C, Spiliotopoulos D, Musco G, Spitaleri A (2015) GMXPBSA 2.1: A GROMACS tool to perform MM/PBSA and computational alanine scanning. *Comput Phys Commun* 186:105–107. <https://doi.org/10.1016/J.CPC.2014.09.010>
  144. Duan L, Liu X, Zhang JZH (2016) Interaction Entropy: A New Paradigm for Highly Efficient and Reliable Computation of Protein–Ligand Binding Free Energy. *J Am Chem Soc* 138:5722–5728. <https://doi.org/10.1021/jacs.6b02682>
  145. Zhu Y, Shmidov Y, Harris EA, et al (2023) Activating hidden signals by mimicking cryptic sites in a synthetic extracellular matrix. *Nat Commun* 14:. <https://doi.org/10.1038/s41467-023-39349-w>
  146. Scoring Tutorial. <https://new.rosettacommons.org/demos/latest/tutorials/scoring/scoring#scoring-in-rosetta>. Accessed 15 Jul 2023
  147. Sengupta D, Prasanna X, Mohole M, Chattopadhyay A (2018) Exploring GPCR-Lipid Interactions by Molecular Dynamics Simulations: Excitements, Challenges, and the Way Forward. *J. Phys. Chem. B* 122:5727–5737
  148. Huang Y, Bharill S, Karandur D, et al (2016) Molecular basis for multimerization in the activation of the epidermal growth factor receptor. *Elife* 5:1–27. <https://doi.org/10.7554/eLife.14107>
  149. Kwilas AR, Donahue RN, Tsang KY, Hodge JW (2015) 乳鼠心肌提取 HHS Public Access. *Cancer Cell* 2:1–17. <https://doi.org/10.1016/j.sbi.2018.07.008.Reprogramming>
  150. Meller A, Ward M, Borowsky J, et al (2023) Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network. *Nat Commun* 14:1–15. <https://doi.org/10.1038/s41467-023-36699-3>
  151. Kwon Y, Shin WH, Ko J, Lee J (2020) AK-score: Accurate protein-ligand binding affinity prediction using an ensemble of 3D-convolutional neural networks. *Int J Mol Sci* 21:1–16. <https://doi.org/10.3390/ijms21228424>
  152. Öztürk H, Özgür A, Ozkirimli E (2018) DeepDTA: Deep drug-target binding affinity prediction. *Bioinformatics* 34:i821–i829. <https://doi.org/10.1093/bioinformatics/bty593>

153. Arul Murugan N, Ruba Priya G, Narahari Sastry G, Markidis S (2022) Artificial intelligence in virtual screening: Models versus experiments. *Drug Discov. Today* 27:1913–1923
154. Maffucci I, Hu X, Fumagalli V, Contini A (2018) An efficient implementation of the Nwat-MMGBSA method to rescore docking results in medium-throughput virtual screenings. *Front Chem* 6:322405. <https://doi.org/10.3389/fchem.2018.00043>

## Copyright forms

<p><b>Citation:</b> Sen N, Kanitkar TR, Roy AA, Soni N, Amritkar K, Supekar S, et al. (2019) Predicting and designing therapeutics against the Nipah virus. PLoS Negl Trop Dis 13(12): e0007419. <a href="https://doi.org/10.1371/journal.pntd.0007419">https://doi.org/10.1371/journal.pntd.0007419</a></p>
<p><b>Editor:</b> Jeanne Salje, University of Oxford, UNITED KINGDOM</p>
<p><b>Received:</b> April 26, 2019; <b>Accepted:</b> November 4, 2019; <b>Published:</b> December 12, 2019</p>
<p><b>Copyright:</b> © 2019 Sen et al. This is an open access article distributed under the terms of the <a href="#">Creative Commons Attribution License</a>, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.</p>
<p><b>Data Availability:</b> All relevant data are within the manuscript and its Supporting Information files. The coordinates of the models of proteins and complexes with the inhibitors is publicly available at <a href="http://cospi.iiserpune.ac.in/Nipah/">http://cospi.iiserpune.ac.in/Nipah/</a></p>
<p><b>Funding:</b> MS Madhusudhan would like to acknowledge the Wellcome Trust-DBT India alliance for a senior fellowship. Neeladri Sen and Sanjana Nair would like to acknowledge CSIR-SPMF for funding. Kaustubh Amritkar would like to acknowledge INSIPRE-SHE fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.</p>
<p><b>Competing interests:</b> The authors have declared that no competing interests exist.</p>

**Keywords:** missense mutation effect prediction, amino acid depth, local environment/cliue, statistical potential, meta predictor

**Citation:** Tan KP, Kanitkar TR, Kwoh CK and Madhusudhan MS (2021) Packpred: Predicting the Functional Effect of Missense Mutations. *Front. Mol. Biosci.* 8:646288. doi: 10.3389/fmolb.2021.646288

**Received:** 25 December 2020; **Accepted:** 19 July 2021;

**Published:** 20 August 2021.

**Edited by:**

[Arun Prasad Pandurangan](#), MRC Laboratory of Molecular Biology (LMB), United Kingdom

**Reviewed by:**

[Wim Vranken](#), Vrije University Brussel, Belgium

[Thiyagarajan S](#), Institute of Bioinformatics and Applied Biotechnology, India

**Copyright** © 2021 Tan, Kanitkar, Kwoh and Madhusudhan. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

**\*Correspondence:** Mallur Srivatsan Madhusudhan, [madhusudhan@iiserpune.ac.in](mailto:madhusudhan@iiserpune.ac.in)

†These authors have contributed equally to this work

**Disclaimer:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.