Computational Methods for Predicting Phenotypic Effects of Mutations Using Large Sequence and Deep Mutational Scan Data

A Thesis

Submitted For the Degree of

Doctor of Philosophy

in the Faculty of Science

by

Sruthi C. K.

Theoretical Sciences Unit
Jawaharlal Nehru Centre for Advanced Scientific Research
Bangalore – 560 064, India

AUGUST 2020

DECLARATION

I hereby declare that the matter embodied in the thesis entitled "**Computational Methods for Predicting Phenotypic Effects of Mutations Using Large Sequence and Deep Mutational Scan Data**" is the result of investigations carried out by me at the Theoretical Sciences Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India under the supervision of Dr. Meher K. Prakash and that it has not been submitted elsewhere for the award of any degree or diploma.

In keeping with the general practice in reporting scientific observations, due acknowledgement has been made whenever the work described is based on the findings of other investigators. Any omission that might have occurred by oversight or error of judgement is regretted.

Sruthi C. K.

CERTIFICATE

I hereby certify that the matter embodied in this thesis entitled "**Computational Methods for Predicting Phenotypic Effects of Mutations Using Large Sequence and Deep Mutational Scan Data**" has been carried out by Ms. Sruthi C. K. at the Theoretical Sciences Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India under my supervision and that it has not been submitted elsewhere for the award of any degree or diploma.

Dr. Meher K. Prakash
(Research Supervisor)

Acknowledgements

When I got selection for Ph.D. in the Theoretical Sciences Unit of JNC, I was excited hoping to work on problems where Quantum Mechanics and Statistical Physics approaches would be used, subjects which were my favourite during masters. But it was on the day of joining that I came to know that I would be working with Dr. Meher K. Prakash, a newly joined faculty then who works on biology related problems. I was really shocked and sad as I always of dreamt of doing problems in Physics. I even thought of changing the advisor, but thanks to the then Chair Prof. Umesh Waghmare and Vasudevan, a senior who suggested to talk to Dr. Meher once before taking such a decision. Following their advice I met Dr. Meher who very nicely explained how basic concepts in Physics and Chemistry can be used to understand interesting phenomena in Biology. That was the beginning of my Ph.D. journey in Computational Biology and I am extremely thankful to my supervisor Dr. Meher K. Praksh for his support and guidance during that journey. He introduced me to a number of exciting questions in Biology as well as a variety of computational tools that can be used to answer those questions. He has always been positive about solving any trouble that we encountered, trying all possibilities, at times pushing me to go beyond my comfort zone and learn new things. He also taught how to stay focussed not getting lost in the learning process. I am very grateful for his wonderful mentorship during which he trained me to ask creative questions, think differently and communicate the findings in a way that can be understood by everyone. He encouraged us to discuss projects among ourselves as well as help each other developing a friendly and helpful research atmosphere in the lab. He was always approachable, making it possible to discuss any problem I faced including those that were unrelated to my research and his suggestions were very valuable. I always enjoyed the long discussions on scientific and non-scientific topics that we used to have in the lab and thank him for all advices that he called “lessons for life”.

I thank all my collaborators Prof. Hemalatha Balaram, Anupam Singh, Malay R Biswal, Ayan Majumder and Sreedevi Padmanabhan for the fruitful interactions. I am very grateful to Prof. Hemalatha Balaram for her constant guidance and for teaching me the basics of Biochemistry. Her enthusiasm and energy during all discussions had always inspired me.

The course work in JNC was a great learning experience and I thank all my course instructors Prof. U. V. Waghmare, Prof. Subir K. Das, Prof. Swapan K. Pati, and Prof. S. M. Sivaprasad, Prof. Kavita Jain and Prof. Hemalatha Balaram.

I acknowledge all the faculties from TSU, Prof. Sobhana Narasimhan, Prof. Srikanth Sastry, Prof. U. V. Waghmare, Prof. Subir K. Das, Prof. Swapan K. Pati, Prof. N. S. Vidyadhiraja and Prof. Kavita Jain for the support during the course of my Ph.D. I also thank the past and present TSU chairmen for their support.

I am very thankful to JNCASR for the financial support as well as the computational facilities. I thank the CCMS administrators for their help in using the cluster. I express my sincere gratitude for all JNC staff for their support and help during my stay at JNC.

All my present and past labmates have been extremely supportive and I extend my heartfelt thanks to all of them. I am very thankful to Anupam, the only labmate I had when I joined, who insisted to work "smartly" rather than hard and taught me the basics of programming and a number of other computational tools. I thank Malay who has always been there whenever I needed help. His support during all stressful periods was invaluable. I also thank him for teaching me violin. I am also grateful to other labmates, Sandhya for the cooking parties, Soumalya and Soutick for the coffee breaks filled with discussions on a wide variety of topics and the “PJs”, Ashlin, Brijesh, Himanshu and Ayan for all the fun we had together in the lab. I thank all summer students Rajbir, Rashi, Roshan, Nithishwer and Ankit for their support.

It is the advices from seniors that help one survive many of the troubles during Ph.D and I was lucky to have a few nice people, Tarak, Priyanka, Koushik, Jiarul, Shubhojit, Saikat, Sona, Ananthu, Vinutha, Anshul and Rajdeep as seniors who had supported me immensely. I thank each of them for their help. Special thanks to Priyanka for the care and love she extended and the cheerful time we had together.

Hostel life indeed was the most memorable part of Ph.D. life. Jayathi, who was a perfect room mate for me became my best friend too and it was through her that I started interacting with others and be friends with three other wonderful people, Shikha, Chaitali and Ekashmi. I am very thankful to all four of them for the food outings, colorful and surprising birthday celebrations and the very long night chats. Spending time with them has always been refreshing and it was their support that helped me overcome many tough periods. I also thank my batchmates Monoj and Sudip for their support.

I express my sincere love and gratitude for all my friends from school to college, who have been very supportive throughout my Ph.D. Special thanks to Monika and Kurias for making the "The Quiet Group" noisy by sharing even the silliest moment of everyday life which made me have the feel of being together and gave me the freedom to disturb them with my stories.

I was fortunate to have my uncle and aunt in Bangalore and hence a second home which I would visit often. They always welcomed me with tasty home food and movie plans. I am very grateful to them for their love and care.

I am very thankful to my sister for her help and guidance throughout my studies as well as PhD.

Finally, I thank my parents and other family members for their love and constant support.

Synopsis

Proteins are biological molecules that perform a number of critical cellular functions. Any variations or mutations in the amino acid sequence of a protein, mostly occurring by random chance, can affect its structure, function or both. In a given cellular context these mutations can have a downstream effect of increasing or reducing the cellular fitness, the latter of which may be compensated by secondary mutations. Thus to predict the disease phenotypes such as Alzheimer's or cancers or antibiotic resistance associated with the mutations in bacteria, or to understand how to engineer proteins, it is important to understand the effects of single and correlated mutations. Acknowledging the importance of the mutational effects, and using the large scale data which has become available through the advancements in the sequencing technologies, in this thesis I raise a few conceptual questions related to how using computational tools mutational effects can be predicted, interpreted, or used for deciphering phenotypic effects.

Chapter 1 presents a brief introduction to proteins, a survey of the different experimental approaches that are used for understanding how proteins function and how computational approaches using molecular dynamics simulations, bioinformatics and machine learning can be used to complement the learnings from the experimental approaches. The remaining chapters in the thesis are organized under two broad umbrellas: correlated mutations are discussed in Part I, and the prediction and interpretation of mutational effects are discussed in Part II.

While mutations occur randomly, proteins in the cells selected for fitness do have mutational changes in various regions that are correlated. These correlations, which may be quantified using alignments of large sequence libraries of homologous proteins, become a way of inferring structural or functional relations among the different amino acids. In Part I of this thesis, we explore the possibility of using correlated mutations in viruses as a way of defining their complexity (**Chapter 2**), a way of capturing the asymmetry in the correlations (**Chapter 3**), and the incompleteness of the important mutations identified using sequence, structure, or dynamics information (**Chapter 4**).

Viruses have high mutation rates and hence evolve rapidly by accumulating mutations that help them in escaping from the host immune response and drugs. However, most

viruses have only around 10 proteins, making amino-acid level mutational patterns the only way of studying the systems-level complexity. In **Chapter 2**, we attempt to define viral genomic complexity using amino acid co-variance networks. By applying the methods of network analysis, we demonstrate the differences in the nature of the network (random versus power-law) in different viruses and also find an interesting relation between the network density and the virus-Richter scale, the only metric to date which quantifies the mortalities due to each infection.

While amino acid co-variance analyses are used to capture structural and functional interactions between amino acids, there has been no measure to capture the asymmetry in the interactions such as when a compensation is required for a mutation. In **Chapter 3**, we introduce one such way of capturing such asymmetric relations using conditional probabilities, and demonstrate examples of how some compensatory effects are captured by it.

In several occasions, such as while studying allosteric effects in proteins, or for identifying hotspot mutations, one uses correlation based methods. However, the analyses are performed using multiple sequences or protein structure or molecular dynamics, a choice mainly driven by one's scientific training. The implicit assumption being that all methods should identify the same amino acids. In **Chapter 4**, we compare the findings from these three approaches and suggest the lack of completeness in any one, and point to the need for complementing the approaches.

Recent developments in massively parallel mutagenesis experiments, also known as deep mutational scans, have increased the number of mutational effects that can be studied by several orders of magnitude. Deep mutational scans have generated a wealth of mutational effect data for a number of proteins, and are heading towards generating hundreds of thousands of double or triple mutants. In Part II of this thesis, we ask several questions around predicting accurately using artificial intelligence (AI) models (**Chapter 5**), classifying using simpler models (**Chapters 6, 7**) and interpreting the contributions of individual factors to every single mutation (**Chapter 8**).

We ask whether computational predictions can be used for reducing and complementing the deep mutational scan experiments. In **Chapter 5**, we ask two different questions, by what degree one may be able to reduce the number of experiments and if it is possible to strategize the design of the mutational studies, such as a random or a site-directed choice of mutations such as an ANH (Alanine, Aspargine, Histidine) scan, in a way to obtain the best possible predictions with the minimal data.

Although the mutational scans have become exhaustive and accurate, both in experiments and predictions, there is still a missing gap between these libraries of data and the qualitative intuitions about what affects a protein's function. In **Chapter 6**, deep mutational data sets are used to quantify the intuitions about the role of parameters such as conservation, solvent accessibility on protein function. We develop rules of thumb for classifying the mutational effects using any of the individual parameters, and demonstrate how the quality of this classification can be enriched by combining multiple parameters.

Continuing on a similar theme of simplified models, we asked whether it is possible to develop human comprehensible Artificial Intelligence (AI) model for the mutational effects prediction. A typical decision tree used for classification can have a depth of 10 to 100. In **Chapter 7**, we ask if the decision tree for mutational effects prediction can be truncated to fit in an A4 size paper (depth of 5) and hence develop a simpler model with comparable accuracy.

The rules of thumb or a truncated decision tree developed although intuitive, compromise on accuracy. AI community is undergoing a philosophical shift, revisiting the trade-off between accuracy and interpretability, and creating models such as SHapley Additive exPlanations (SHAP) to make the AI predictions interpretable. In **Chapter 8**, we use SHAP to make a first attempt to build interpretable models in the field of mutational effects prediction. Using the contributions of individual factors for every single prediction, we could extract cleaner correlation patterns between the mutational effects and individual variables, that are otherwise hidden in the multi-variable dependence.

Another question we ask with the deep mutational scan libraries, in **Chapter 9**, is if there is an alternative reason behind the codon bias, beyond the tRNA availability, mRNA toxicity. We explore how the choice of the codons can reduce the potentially deleterious mutational effects. The thesis thus takes advantage of the large scale mutational data that is available, uses the developments in network theory, AI to predict or interpret the phenotypic effects arising from the mutations.

List of Publications

1. “Amino acid impact factor”. Sruthi, C.K. and Prakash, M., **2018**. *PloS one*, 13(6).

2. “Statistical characteristics of amino acid covariance as possible descriptors of viral genomic complexity”. Sruthi, C.K. and Prakash, M.K., **2019**. *Scientific Reports*, 9(1), pp.1-12.

3. “Deep2Full: Evaluating strategies for selecting the minimal mutational experiments for optimal computational predictions of deep mutational scan outcomes”. Sruthi, C.K. and Prakash, M., **2020**. *PloS one*, 15(1), p.e0227621.

4. “Towards developing intuitive rules for protein variant effect prediction using deep mutational scan data”. Sruthi, C. K., Balaram, H. and Prakash, M. K., *(under review)*

5. “Interpreting mutational effects predictions, one substitution at a time”. Sruthi, C. K., and Prakash, M. K. **(2019)**. bioRxiv, 867812, doi:https://doi.org/10.1101/867812 *(to be submitted)*

6. “An interpretable prediction of mutational effects using a compact A4 size decision tree”. Sruthi, C. K., Biswal, M. R. and Prakash, M. K. *(under review)*

7. “Correlations from sequence, structure and dynamics are complementary rather than synonymous”. Sruthi, C. K., Singh, A., Balaram, H., and Prakash, M. *(manuscript under preparation).*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Proteins

1.1.1 Protein Structure

Proteins are very important biological macromolecules that are involved in many cellular processes. Transport molecules such as haemoglobin, antibodies which are related to immune response, enzymes which catalyze chemical reactions, and structural matrices such as keratin or collagen are all proteins. In an apparently striking recursive role, all proteins are synthesized by other special proteins called ribosomes.[1] Thus any quest for understanding the basic cellular or disease biology mostly narrows the search to activity of some proteins or their failure. Whether it is non-communicable diseases such as cancers or Alzheimer's or communicable diseases with bacterial or viral infections, the fundamental interest is always in knowing what went wrong with the most important proteins in healthy cells or how to block the bacterial or viral proteins from performing their expected functions.[2–4]

Chemically speaking, proteins are polypeptide chains, linear polymers of amino acids connected by peptide bonds.[5] They are synthesized by ribosomes, with combinations of the 20 naturally occurring amino acids, appearing in a unique sequence. The linear polypeptide chain, which is also known as the primary structure of the protein, undergoes further physical transformations and levels of organization before it becomes functional.[6] Driven by non-bonded interactions such as hydrogen bonds or salt bridges among the amino acids which are either near or distal in sequence, proteins achieve secondary structures such as helices, β-sheets, turns or coils, which are further organized into a tertiary folded structure.[6] The ability to form different secondary structures varies for the side chains of different amino acids.[7] An even further organization into a quaternary structure happens with proteins which are multimeric in nature.[8]

The sequence hypothesis categorically states that the three dimensional structure and the function of the protein are both uniquely determined by its sequence.[9] While the hypothesis is empirically true, several aspects of the sequence-structure-function relation in the proteins are not understood. Three dimensional protein structures that are formed by a hydrophobic collapse are delicately balanced by the loss of entropy and gain of enthalpy.[10, 11] From this delicate balance, which is only a marginal stability of a few *kcal/mol* for most of the proteins, it has been difficult to predict the unique structure of the protein. A well defined structure is very crucial for the proper functioning of a protein,[6] and its knowledge is required for understanding how a specific protein works or how to design drugs that target it.[12, 13]

The most commonly used method to determine the three-dimensional structure of proteins is X-ray crystallography, where the purified and crystallized protein is subjected to X-ray beam and the resulting diffraction pattern is analyzed to infer the structure.[5] Though the structure can be obtained at atomic detail, the requirement that the protein needs to be crystallized, which is a difficult task, limits the applicability of the method. Further, it is difficult to infer the structure of the flexible regions of the proteins using X-ray crystallography, as they appear to be the missing regions in the crystal structure. Whether the crystal structure does have a correspondence with the protein structure in solution while performing its function is an additional question that is sometimes cast on the X-ray structures. Electron microscopy techniques are also being used for obtaining domain level information in very large protein assemblies or of entire virus.[14, 15] For atomistic mechanisms and interpretations, these analyses should be combined with data on finer scale structures.[16]

1.1.2 Protein Dynamics

However, protein structure is only one of the components contributing to its function. Dynamics of the protein is critical for resulting in the functional effects.[17, 18] The simplest illustration of dynamics is with enzymes, where parts of the protein (typically referred to as the lid) have to move to create space for the substrate(s) to enter the catalytic region of the enzyme.[19] While the actual dynamics itself can not be captured, cryo-electron microscopy does reflect the dynamic heterogeneity of the protein, providing a snapshot of the configurations that are assumed by the protein, some with higher probability and others with much lower probability.[20]

Tracking the dynamics of the proteins thus maps the transitions across these different configurational or functional states. One of the extremes of the dynamical transitions happens soon after the synthesis of the protein, where the long unstructured polypeptide chain folds into the native structure.[21, 22] One of the early interests of the physicists who were fascinated by proteins was the modelling of their folding landscapes. It was mainly motivated by the puzzle surrounding the quick rates of protein folding, despite a feared disparity as raised by Levinthal's paradox. Several models such as the folding-funnel concept addressed some of these issues.[23, 24] Much of the functionally interesting structural transitions are less dramatic, with some of them requiring as little as an Angstrom of movement of the appropriate amino acids.[25] The functionally relevant dynamics thus happen in a smaller subset, conformational subspace of all possible dynamical transitions.

Physiological length and time scales

Exploring the biologically interesting dynamical processes in proteins span across different time scales, from a few femtoseconds to seconds. Interestingly all of these are spontaneous and driven by the thermal motions of the atoms, which suggests that chemical and physical intuitions can be useful for studying the biological phenomena. For example, vibrational motions are in pico-second time scale where as large conformational changes like domain motions happen in milliseconds timescale while the folding and unfolding events occur on a few minutes to a few days respectively, depending upon the stability of the proteins. On the one extreme of studying the dynamics of proteins, popular techniques such as Fluorescence resonance energy transfer (FRET) across the fluorophores at different locations in the protein are very helpful to capture distance changes on the order of 20-80 Å. Vibrational spectroscopic studies, on the other hand, probe the local vibrational modes of the protein on the scale of a few picoseconds using infrared techniques, for example. Finding the appropriate tools in the scales of length (0.5-5 Å) and times (µs-s), where the protein is mostly in its native state and performing biological activities is more complicated than working with either of the extremes.

Fluorescence techniques are a major source of dynamic information about proteins. They make use of the fluorescence property of the aromatic groups of the amino acids - tryptophan, tyrosine and phenylalanine - by having them as the probes to study protein conformation, dynamics, and intermolecular interactions. Given the timescale separation from the radiative fluorescence decay, these methods are ideal for monitoring dynamics on microseconds or longer timescales. However, the information one gathers from these studies is coarse grained and is mostly limited to understanding whether the excited molecule can undergo a radiative transition or it is hindered by the local environment of the excited amino acid. A systematic replacement of different amino acids with tryptophans can provide the details of the conformational changes at different positions of interest.[26] The implicit assumption in this approach is that the structure or function of the protein is not compromised by this replacement of amino acids, which is not true in many cases.

Nuclear magnetic resonance (NMR) in contrast to a crystallized structure with a tight packing is used for resolving the solution structure that is most relevant to the cellular context. Relaxation of ^{15}N nuclei of backbone is affected by the dynamics of the protein and measuring this relaxation time helps in identifying the flexible regions of the protein.[27] Deuterium does not have a peak in the NMR spectrum, and this fact is exploited by probing the hydrogen exchange reaction (NH →ND) of the backbone amide when the solvent is heavy water. The hydrogen of the backbone amide gets exchanged upon exposure to deuterated solvent, resulting in the disappearance of the corresponding peak. The disappearance of the peaks thus reflect the conformational changes that make residues solvent exposed over time.[27] Thus, NMR can determine the solution structure of small proteins and hence makes it possible to obtain structural and dynamical insights into the functional form of the proteins. The technique is especially suited to the functional timescales of milliseconds to seconds which are of functional relevance. However, these techniques are usually limited to smaller proteins with 50-100 amino acids.

Computational methods: Molecular dynamics

While dynamics of the proteins driven by thermodynamic forces are what cause the proteins to behave as they should to be performing the functions in living cells, these microscopic details across the length and timescales of interest are not easy to be captured by any of the experimental techniques. Molecular dynamics (MD), a method of computer simulation, has emerged as a complementary approach to the experimental techniques.[28, 29] Classical MD simulations are based on evolving the trajectories of atoms in the molecular systems, with the parameterized interaction potentials between bonded and non-bonded atoms (force-fields) according to Newton's laws of motion. Specifically for biomolecular systems, the MD simulation algorithms can be used for obtaining a few millisecond long trajectories on a few million atoms corresponding to a few hundreds of amino acids (along with other accompanying atoms from water, membranes, etc). The force-fields which are based on quantum mechanical calculations, are optimized for the native structure of the protein. Hence, the MD simulation trajectories have been able to capture several interesting aspects of protein dynamics and function, starting from the fleeting hydrogen bonds within the protein with the solvent to the conformational changes of protein domains, to functional motions such as opening of the enzyme lids. But arriving at the native fold and following the conformational dynamics within the native fold are both different and difficult challenges and one is mostly limited to working with native conformation obtained from X-ray crystallography or NMR, to study the dynamics that happens around this configuration. Advanced sampling methods such as replica exchange molecular dynamics (REMD),[30] umbrella sampling,[31] Metadynamics[32] have enhanced the sampling of several large scale movements, but despite these advances MD is still far from being the complete tool for microscopic analysis on biological systems where a clear separation of degrees of freedom and timescales is not possible.

Computational models: Bioinformatics

As the physics based simulation models and experiments suffer from their inherent limitations, information based models are gaining prominence. The basis of this focus on bioinformatic tools, is on several grounds – the sequence hypothesis asserts that the sequence information is sufficient for understanding the structure and function (and implicitly the dynamics) of the protein, because of the advances in the sequencing technologies an increasing number of sequences of proteins that are structurally or functionally related are becoming available, machine learning algorithms with implicit rules are able to predict patterns which are otherwise hard to derive from explicit rules that are based on physical parameters and forces. As such bioinformatic tools are becoming increasingly relevant to predict protein structure,[33] and function.[34, 35]

Using a multiple sequence alignment (MSA) of sequences of DNA, RNA, or amino acid codes across different proteins or their variants could be analyzed for understanding functional, structural, or evolutionary relationships among these sequences. The information from the N × L matrix of N sequences all of length L after inserting the required gaps ("-")to align the different sequences becomes the starting point of various analyses. The simplest one being the variability across any column of this matrix, which reflects the lack of conservation of the amino acid and how it is tolerated across homologous positions. The next level of consideration is the groups of contiguous regions in this aligned sequence which can characterize protein families, identify shared regions of homology from the multiple sequence alignments. Consensus sequences formed by identifying the most occurring amino acid at each of the L positions can help to develop a sequence "finger print" which allows the identification of members of distantly related protein family (motifs). While the variability of amino acids across species is not too uncommon, sometimes a minor change such as the change of a single amino acid (mutation) within the same species can lead to a loss of the protein's structure or function.

The reason for the persistent structure or function with multiple changes or mutations in the proteins has been typically attributed to compensatory effects[36] of either the structurally neighboring amino acids or amino acids which are distal in sequence and structure but yet affect through allosteric effects.[37] A few frameworks for obtaining the relations between distal amino acids have been proposed based on the simultaneous changes that happen while considering pairs of amino acids in the multiple sequences. Co-evolution,[38] is a statistical measure of how pairs of amino acids change relative to their respective consensus positions. With no immediate causal inference, or a sense of spatial distal distance, these co-evolutionary relations when they are strong, suggest a statistical relation of simultaneously changing amino acids. Starting from the co-evolution as a basis, and by applying the appropriate eliminating higher order effects from the co-evolution matrices, other formalisms direct coupling analysis (DCA)[39] and statistical coupling analysis (SCA)[40] have been developed. These different analyses were meant to capture the distance metrics between co-evolving amino acids with the intention of *de novo* structure prediction[33] and the functional demarcations and correlations across the different contiguous regions, also known as "sectors" respectively.[40]

However, as with all experimental techniques, and molecular dynamics simulations, there are limitations to the sequence based methods. The results may be biased by the numbers and types of sequences that are chosen. The de novo structure prediction has been successful on smaller proteins, and of course the while an amino acid may be deemed important, it is not easy to understand whether the dynamical nature of the protein confers this role. The methods using structure, dynamics and sequence have been evolving independently, shining light on complementary aspects of protein function.

1.1.3 Mutations

In most practical concerns of understanding basic cellular or disease biology, what is interesting in several occasions is something that is more fine grained than the details of the large-scale long-timescale unfolded to folded transformation, and simpler than determination of the complete three dimensional structure of the protein or the conformational changes that happen on milliseconds timescales. Most common mutations in the DNA are at the level of a single nucleotide polymorphism (SNP). These SNPs are referred to as synonymous or non-synonymous depending on whether they result in a change in the amino acid they code for or not. Deletion or insertion of base pairs can either lead to deletion or insertion of amino acids or even more complicated changes such as frame shift mutations.[1] It is empirically believed that the the mutation rate correlates inversely with the length of the genome.[41] As such, most common changes in the genetic material, implicated in diseases like cancer,[42] sickle-cell anemia[43] or in studying the development of drug resistance in bacteria can mostly be traced to SNPs.[44, 45]

Understanding the effects of SNPs, which result in at most a change of a single amino acid in a protein of a few hundred amino acids, on the structure or dynamics of proteins are extremely important. At a simple intuitive level, the effect seems perturbative, as a small perturbation on top of a large background. However, this is far from the truth, as some of the single mutations can cause structural instability or/and functional impairment, and others leave the function of proteins unaltered.

Mutational effects

There are several biochemical intuitions about amino acid substitutions. For example, proline and glycine are called helix breakers[6] as these amino acids cannot form the hydrogen bonds to maintain the helical structure. Substitutions to these amino acids can disturb the secondary structure of the proteins. Mutations occurring at the sites which are buried are more likely to lead to a loss of structure or function. Substitutions of amino acids which lead to a change of charge type or a decrease in volume can be deleterious. However, proteins are complex and these intuitions do not always hold. Thus, understanding the effects of mutations *in vitro* and *in cellulo* requires more detailed experimental studies complemented by theoretical models.

Mutagenesis studies

Mutagenesis experiments which introduce random or systematic substitutions of amino acids in a protein sequence were used to assess the functional consequences of mutations. Initial studies had access only to natural variants and mutations that occur spontaneously.[46]. Later the use of X-rays[47] and chemical mutagens[48] by which mutations could be induced randomly revolutionized the field. But the major limitation of this classical in-vivo approach was that it required a specific phenotype to select the variant of interest from the pool of thousands of mutants. Development of transposon mutagenesis technique[49, 50] where natural mobile segments of DNA could be transferred to the genome of interest made it possible to make a single insertion mutation in bacterial genomes.[51]

However, with the advent of newer *in-vitro* mutagenesis technologies such as random, site directed mutagenesis, any desired mutation could be performed. Error-prone polymerase chain reaction (epPCR)[52] is commonly used to induce mutations randomly along the DNA sequence. But it is not possible to get all possible amino acid substitutions through this method. On the contrary, oligonucleotide primers can be used to introduce specific mutation at a desired location. Unlike random mutagenesis this method of site-directed mutagenesis[53] can generate all mutants, but is expensive and time consuming. Together the two techniques have empowered biochemists to perform mutations for understanding the functional effects of mutations. Even interestingly, the unexpected changes in kinetics of folding or catalysis due to mutations have revealed how the different amino acids are involved in the transition states of these different processes.[5] One of the techniques which evolved as a systematic way of evaluating the roles of different amino acids is alanine scan.[54] Factors like non-bulky side chain, retains main-chain conformation and charge neutrality make alanine a natural choice for this substitution scan. In alanine scan mutagenesis[54] each amino acid in the protein is replaced by alanine, one at a time, to study the importance of an amino acid in function or maintaining the structural stability of the protein.

Deep mutational scans

Though generating mutant libraries were feasible with these methods, selection of mutants required developments in the sequencing technologies. With the introduction of high-throughput screening, large libraries of variants could be screened. A new paradigm in mutational scans, deep mutational scan (DMS), has emerged.[55] DMS involves probing the functional/structural consequences of all possible single point mutations in a protein, considering one substitution at a time. The simplest deep mutational scan is a set of thousands of single point mutations on one single protein.[55] Newer studies[56–58] are pushing the envelope of the mutations to simultaneous multiple mutations, increasing the number of substitutions variants studied to a few hundred thousands. The two positive aspects of the DMS are that the problem of protein purification is averted and the phenotypic consequences of the mutations which are the downstream effects of the protein mutations can be studied. It uses any of the mutagenesis techniques along with high-throughput screening to select for mutants of desired phenotype. The change in population size of each variant before and after the selection condition which is relevant for the function of the protein under study quantifies the effect of mutation on the fitness of the organism. Figure 1.1 illustrates this concept pictorially as well as shows how the deep mutational scanning data is typically visualized.

Phenotypic effects

If predicting the structure or dynamics of proteins, or their subtle changes due to mutations are difficult, predicting the downstream effects of a protein modification at a cellular level (phenotypic effects) are even more complicated. However, these are the studies that are most relevant for understanding diseases or devising possible cures. Molecular dynamics studies, which are with atomistic details, are typically repeated for every mutation that needs to be studied. This trivial repetition of calculations does not scale well computationally when several mutations are physiologically interesting. This is in reality a limitation of the, otherwise very attractive, models which are based on physical intuitions.

AI based predictions

As the emergent phenomena appearing from the interactions of hundreds of thousands of degrees of freedom within a protein and subsequent protein-protein interactions become

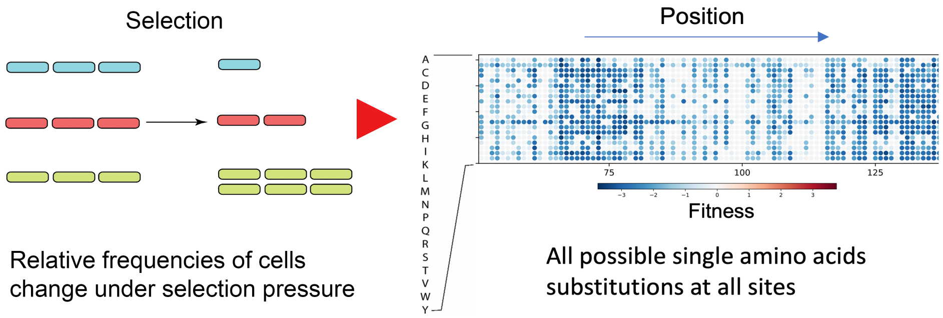
Selection
Relative frequencies of cells change under selection pressure
Position
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V
W
Y
75
100
125
−3
−2
−1
0
1
2
3
Fitness
All possible single amino acids substitutions at all sites

Figure 1.1: Schematic representation of deep mutational scanning experiments: Deep mutational scanning is a new way of performing mutational scanning simultaneously on thousands of variants of a protein. Variants of the protein of interest are expressed in/transformed to bacterial cells and the frequency of cells with different variants is monitored as the cellular function of the protein is challenged. The cells with functionally advantageous variants increase in frequency and the changes in the relative frequencies after the selection process are used to infer the effect of mutation on cellular fitness. The variant effect scores obtained in this way for all possible single amino acid substitutions in a protein are typically represented as a colormap as shown on the right.

harder to predict, one resorts to the implicit models based on artificial intelligence (AI) or machine learning (ML). AI based techniques have gained lot of interest in the recent past for modeling many complex problems in biology[59] like predicting structure and function of proteins. For example, in what is called supervised learning, the AI model is trained on a sample set of data ($\mathbf{X}$,$\mathbf{y}$), with input vector $\mathbf{X}$ and output $\mathbf{y}$. An example of ($\mathbf{X}$,$\mathbf{y}$) in the biological contexts we just described is a set of descriptive parameters $\mathbf{X}$ which quantify the nature of the mutation and $\mathbf{y}$ can be the change in the cellular fitness. The size of the data set thus goes as $m \times (n + 1)$, where for each of the m different experiments or trials or in the example mentioned the number of mutations, one has the data for the n descriptive parameters and the observed output. In a traditional regression approaches, the emphasis is on obtaining the functional relation $y = f(X)$. However with complicated systems, this functional relation may be non-trivial. Further in most practical scenarios, one may require reliable predictions even if deriving an explicit mathematical model relating them is not possible.

An example of a supervised learning AI model is Neural Networks (NN) model.[60] The model mimics the way human brain learns and recognizes patterns and has a layered structure similar to the connected neurons in our brain. A typical neural network architecture has one input layer, one or more hidden layers and one output layer as shown in Figure 1.2. Input layer is the first layer where the features or variables used for prediction are fed, and the output layer is the last layer. The layer(s) in between the

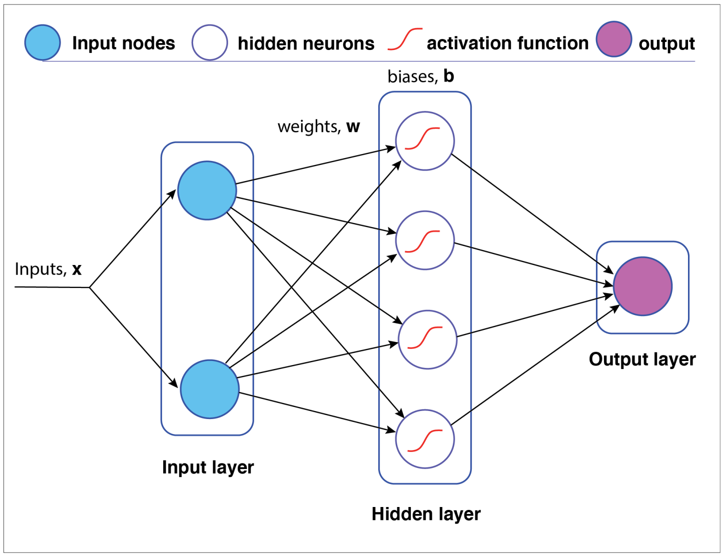

Figure 1.2: Schematic of a shallow neural network. The architecture of a shallow neural network with 2 inputs, 1 hidden layer with 4 neurons and 1 output is shown. w's are the weights and b's are the biases. At a hidden neuron j, the inputs are processed as $g(z_j)$ where g is the activation function and $z_j = \sum(w_{ij}\, x_i) + b_j$. w_{ij} is the weight of i^{th} input at j^{th} hidden neuron and b_j is the bias at neuron j. The output of the hidden neuron is passed to the next layer. One of the common choices for g, a sigmoidal function is also shown inside the hidden neurons.

input and output layers are known as the hidden layer(s) and can be more than one. The nodes in each layer are called neurons. The non-linear transformation is performed in the hidden neuron using an activation function and the output of each layer is the input for the next layer. Training of the neural network, involves the optimization of the model parameters weights, W's and biases, b's. It is possible that the network generated after training predicts the data used for training very well, but cannot predict any unseen data. This is termed as over-training and there are different methods to prevent over-training. In early stopping method, an additional data set called the validation set is used for checking the quality of prediction on untrained data, there by exercising a check for over-training before proceeding to the final predictions on the test data set. Validation set does not directly influence the tuning of weights and biases, but decides when the training has to be stopped. As the error in prediction of training reduces upon training, error in validation set also will reduce upto a certain number of iterations after which it starts increasing. The training process is then stopped.

Decision trees are another AI based approach for supervised learning and can be used for both regression and classification.[61] It works by splitting the training set data samples into two subsets based on any of the input feature so as to have lesser variability for the dependent variable within each subset. Each of these subsets are further divided into two and this splitting is continued until each subset achieves a desired homogeneity for the dependent variable or further division does not improve the homogeneity. This approach of generating decision tree is called recursive binary splitting. The typical architecture of a decision tree with a depth of 2 grown following this procedure is shown in Figure 1.3. The cost function or the measure of inhomogeneity which is minimized at each split can be different, for example in classification Gini impurity defined as $1 - \sum_{i=1}^{N} p_i^2$ where p_i is the fraction of occurrence of i^{th} class at a given node. The weighted sum of the Gini impurity of the subnodes is calculated for division based on each input feature and the one which brings highest reduction in the impurity compared to the parent node is chosen. Entropy, defined as $\sum_{i=1}^{N} -p_i \, log_2 \, p_i$ is also used as cost function. For regression, the variance of the dependent variable is used as cost function. One of the advantages of the decision trees are that it is simple, intuitive and less abstract as compared to other AI approaches. On the other hand it suffers from the problem of overfitting and combining predictions of an ensemble of decision trees is a widely used approach to tackle overfitting, such as the random forest method[62]. In random forest, decision trees are created by training on randomly selected N data points from a trainimg set of size N, but the samples being chosen with replacement which means the same sample can occur multiple times. Trees created in this way are different as they are trained on different data sets. To make the multiple trees even more uncorrelated, in random forest only a subset of features are considered for split at a given node. For classification the class assigned by majority of the trees in the ensemble is the predicted class and for regression the average of prediction from all trees is the final prediction.

Supervised learning methods, where a part of the data is used for training the model, are used in AI for either for making quantitative predictions of the output, $\mathbf{y}$, which is a continuous variable or for classifying the data based on the features, where the output $\mathbf{y}$ is now a categorical variable. Unsupervised learning involves learning from the input vector $\mathbf{X}$ alone. The output is not known and the model adapts as new data becomes available. Unsupervised learning helps in understanding the structure in the data, minimal dimensions that are required for describing the system, clustering to identify groups and identify association rules between groups.

AI for mutational effects

Machine learning algorthims have been used for predicting the functional effects of mutations on proteins. By training on mutational data gathered over decades from across different studies, using biophysical or biochemical intuitions as descriptive parameters, the several models for predicting the effects of mutations emerged. SIFT[63] and Evolutionary

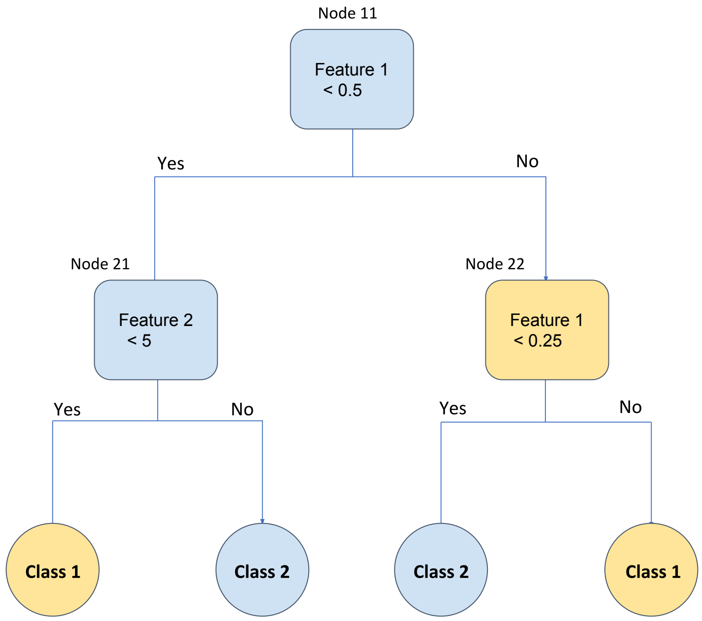

Figure 1.3: Schematic of a decision tree. Illustration of the architecture of a trained decision tree with a depth of 2 for classification is shown. The rectangles and circles represent the nodes of the tree. The top most node is called root node and the 4 nodes in the lowest level that do not split are called leaf nodes. The nodes are colored according to the majority class at each node. The data points are fed to the root node (node 11) of the decision tree, and the ones that satisfy the condition specified at that node will reach node 21 and others at node 22. At node 21 the data points are subdivided based on the value of feature 2 and at node 22 based on the value of feature 1. The class of each data point is then predicted depending on the leaf node they end up in.

action use evolutionary information for the prediction, SNAP2[64] different sequence-related information, Polyphen2 both sequence and structure information and SNPs3d[65] - structural. SNAP2[64] and Polyphen2[66] are supervised methods, both based on machine learning techniques. SNAP2 is trained on an extensive data set of Protein Mutant Database[67] and Polyphen2 on human Mendelian disease-associated and neutral variants. All of these except SNAP2 are classifiers classifying mutation as neutral and non-neutral. SNAP2 score quantifies the mutational effect. Quantitative prediction of mutational effects has been attempted[68, 69] from the pair-wise co-variation of amino acids (EVmutation). Later accounting for multi-site interactions, DeepSequence,[69] improved the predictions further. DeepSequence has been found to outperform other methods in predicting the functional effects of mutations.[70] However, like all other models, these predictions are far from being complete. For example, how a mutation affects the cellular fitness may vary qualitatively depending upon the drug used against the growth of these cells *i.e.*, the functional effects of mutations are context dependent. This brings the discussion back to the molecular nature of the mutations or the factors that interact with the proteins.

1.1.4 Challenges to be addressed

The deep mutational scan experiments are advancing to unravel the effects of simultaneous double or triple mutations with hundreds of thousands of studies on the same protein, simultaneously cryo-electron microscopy based methods are revealing heterogenous structures of proteins including how the interaction complexes are formed. Most theorists attempt to understand the molecular details such as the mechanisms of diseases, suggest newer modifications for example like engineering proteins, or to reduce the experimentation that is required. However, in the steps towards understanding the function of proteins or its modifications, the theoretical models, while very advanced are still catching up with the experiments. The timescales involved, complex non-perturbative nature of interactions and the desire for understanding a molecular level picture leaves no one tool whether it is molecular dynamics or machine learning as the perfect tool for understanding the function of proteins. Obtaining complementary learnings from the protein sequences, structures and dynamics to predict functions seems almost unavoidable. When a wholistic understanding of the proteins is required to develop an understanding into the most critical functions of the cells, leaving out any one of the approaches only compromises the goal.

There are several challenges faced by the biologists today. Bacteria have been developing antibiotic resistance. For example, newer strains of proteins called β-lactamase which is instrumental in the antibiotic resistance of *E. coli* or *K. pneumonia* are developing.

One such example is the New Delhi β-lactamase, which is a resistant strain that is now globally spread. Viruses with about 10 proteins, have a genome size that is at least 1000 times smaller than the bacterial genomes. Although the genome is much simpler, it is harder to treat many viral infections - HIV or avian influenza - for example. Some how the rate of mutation, which is inversely proportional to the genome size, compensates for the fewer numbers of proteins, and confers the fitness advantage which is otherwise surprising. Disease models, which study the evolution or epidemiology of the infections still fail to make apparently simple predictions such as which is the strain of flu that is next going to infect a country.

Several molecular level and theoretical questions arise in the light of the challenges faced by communicable or non-communicable diseases. How do mutations occur, which are the mutations are deleterious or advantageous? Are the origins of the effects of mutations structural or dynamical in nature? If a mutation is deleterious, should it be compensated by others? Is it easy to assess how many such compensatory mutations may be expected to occur in nature? How do different amino acid substitutions affect the function of the proteins? Can the solubility which is a very important criterion for the heterologous expression of proteins be tuned with amino acid substitutions without significantly compromising their function? These are some of the questions we explore in the thesis, motivated by experiments, combining the sequence-structure-dynamical information as much as possible.

Bibliography

[1] H. Lodish, A. Berk, C. A. Kaiser, M. Krieger, M. P. Scott, A. Bretscher, H. Ploegh, P. Matsudaira, *et al.*, *Molecular cell biology.* Macmillan, 2008.

[2] S. Stefl, H. Nishi, M. Petukh, A. R. Panchenko, and E. Alexov, “Molecular mechanisms of disease-causing missense mutations,” *Journal of molecular biology*, vol. 425, no. 21, pp. 3919–3936, 2013.

[3] S. Hasan, S. Daugelat, P. S. Rao, and M. Schreiber, “Prioritizing genomic drug targets in pathogens: application to mycobacterium tuberculosis,” *PLoS computational biology*, vol. 2, no. 6, 2006.

[4] S. K. Chanumolu, C. Rout, and R. S. Chauhan, “Unidrug-target: a computational tool to identify unique drug targets in pathogenic bacteria,” *PloS one*, vol. 7, no. 3, 2012.

[5] A. Fersht *et al.*, *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding.* Macmillan, 1999.

[6] C.-I. Brändén and J. Tooze, *Introduction to protein structure.* Taylor & Francis, 1999.

[7] P. Y. Chou and G. D. Fasman, "Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins," *Biochemistry*, vol. 13, no. 2, pp. 211–222, 1974.

[8] D. Whitford, *Proteins: structure and function.* John Wiley & Sons, 2013.

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

[10] D. Voet and J. G. Voet, "Biochemistry, 4-th edition," *NewYork: John Wiley& SonsInc*, vol. 492, 2011.

[11] C. M. Dobson, A. Šali, and M. Karplus, "Protein folding: a perspective from theory and experiment," *Angewandte Chemie International Edition*, vol. 37, no. 7, pp. 868–893, 1998.

[12] C. A. Orengo, A. E. Todd, and J. M. Thornton, "From protein structure to function," *Current opinion in structural biology*, vol. 9, no. 3, pp. 374–382, 1999.

[13] A. C. Anderson, "The process of structure-based drug design," *Chemistry & biology*, vol. 10, no. 9, pp. 787–797, 2003.

[14] J. Frank, "Single-particle reconstruction of biological molecules–story in a sample (nobel lecture)," *Angewandte Chemie International Edition*, vol. 57, no. 34, pp. 10826–10841, 2018.

[15] W. Jiang and L. Tang, "Atomic cryo-em structures of viruses," *Current opinion in structural biology*, vol. 46, pp. 122–129, 2017.

[16] W. Chiu, M. L. Baker, W. Jiang, M. Dougherty, and M. F. Schmid, "Electron cryomicroscopy of biological machines at subnanometer resolution," *Structure*, vol. 13, no. 3, pp. 363–372, 2005.

[17] N. R. Latorraca, A. Venkatakrishnan, and R. O. Dror, "Gpcr dynamics: structures in motion," *Chemical reviews*, vol. 117, no. 1, pp. 139–155, 2016.

[18] T. Saleh and C. G. Kalodimos, "Enzymes at work are enzymes in motion," *Science*, vol. 355, no. 6322, pp. 247–248, 2017.

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

[20] K. Murata and M. Wolf, “Cryo-electron microscopy for structural analysis of dynamic biological macromolecules,” *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1862, no. 2, pp. 324–334, 2018.

[21] C. M. Dobson and M. Karplus, “The fundamentals of protein folding: bringing together theory and experiment,” *Current opinion in structural biology*, vol. 9, no. 1, pp. 92–101, 1999.

[22] K. A. Dill and J. L. MacCallum, “The protein-folding problem, 50 years on,” *science*, vol. 338, no. 6110, pp. 1042–1046, 2012.

[23] P. E. Leopold, M. Montal, and J. N. Onuchic, “Protein folding funnels: a kinetic approach to the sequence-structure relationship.,” *Proceedings of the National Academy of Sciences*, vol. 89, no. 18, pp. 8721–8725, 1992.

[24] J. N. Onuchic, N. D. Socci, Z. Luthey-Schulten, and P. G. Wolynes, “Protein folding funnels: the nature of the transition state ensemble,” *Folding and Design*, vol. 1, no. 6, pp. 441–450, 1996.

[25] K. Teilum, J. G. Olsen, and B. B. Kragelund, “Functional aspects of protein flexibility,” *Cellular and Molecular Life Sciences*, vol. 66, no. 14, p. 2231, 2009.

[26] C. A. Royer, “Probing protein folding and conformational transitions with fluorescence,” *Chemical reviews*, vol. 106, no. 5, pp. 1769–1784, 2006.

[27] I. A. Kaltashov and S. J. Eyles, *Mass spectrometry in biophysics: conformation and dynamics of biomolecules*, vol. 12. John Wiley & Sons, 2005.

[28] M. Karplus and J. A. McCammon, “Molecular dynamics simulations of biomolecules,” *Nature Structural & Molecular Biology*, vol. 9, no. 9, p. 646, 2002.

[29] M. Karplus and J. Kuriyan, “Molecular dynamics and protein function,” *Proceedings of the National Academy of Sciences*, vol. 102, no. 19, pp. 6679–6685, 2005.

[30] Y. Sugita and Y. Okamoto, “Replica-exchange molecular dynamics method for protein folding,” *Chemical physics letters*, vol. 314, no. 1-2, pp. 141–151, 1999.

[31] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, “The weighted histogram analysis method for free-energy calculations on biomolecules.

i. the method," *Journal of computational chemistry*, vol. 13, no. 8, pp. 1011–1021, 1992.

[32] A. Laio and M. Parrinello, "Escaping free-energy minima," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12562–12566, 2002.

[33] D. S. Marks, T. A. Hopf, and C. Sander, "Protein structure prediction from sequence variation," *Nature biotechnology*, vol. 30, no. 11, p. 1072, 2012.

[34] M. M. Gromiha and Y.-Y. Ou, "Bioinformatics approaches for functional annotation of membrane proteins," *Briefings in bioinformatics*, vol. 15, no. 2, pp. 155–168, 2014.

[35] D. W. Buchan and D. T. Jones, "The psipred protein analysis workbench: 20 years on," *Nucleic acids research*, vol. 47, no. W1, pp. W402–W407, 2019.

[36] M. S. Breen, C. Kemena, P. K. Vlasov, C. Notredame, and F. A. Kondrashov, "Epistasis as the primary factor in molecular evolution," *Nature*, vol. 490, no. 7421, pp. 535–538, 2012.

[37] D. N. Ivankov, A. V. Finkelstein, and F. A. Kondrashov, "A structural perspective of compensatory evolution," *Current opinion in structural biology*, vol. 26, pp. 104–112, 2014.

[38] S. W. Lockless and R. Ranganathan, "Evolutionarily conserved pathways of energetic connectivity in protein families," *Science*, vol. 286, no. 5438, pp. 295–299, 1999.

[39] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, "Identification of direct residue contacts in protein–protein interaction by message passing," *Proc. Natl. Acad. Sci. USA*, vol. 106, no. 1, pp. 67–72, 2009.

[40] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, "Protein sectors: evolutionary units of three-dimensional structure," *Cell*, vol. 138, no. 4, pp. 774–786, 2009.

[41] J. W. Drake, "A constant rate of spontaneous mutation in dna-based microbes.," *Proceedings of the National Academy of Sciences*, vol. 88, no. 16, pp. 7160–7164, 1991.

[42] N. Deng, H. Zhou, H. Fan, and Y. Yuan, "Single nucleotide polymorphisms and cancer susceptibility," *Oncotarget*, vol. 8, no. 66, p. 110635, 2017.

[43] K. Y. Fertrin and F. F. Costa, "Genomic polymorphisms in sickle cell disease: implications for clinical diversity and treatment," *Expert review of hematology*, vol. 3, no. 4, pp. 443–458, 2010.

[44] C. Chewapreecha, P. Marttinen, N. J. Croucher, S. J. Salter, S. R. Harris, A. E. Mather, W. P. Hanage, D. Goldblatt, F. H. Nosten, C. Turner, *et al.*, "Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes," *PLoS genetics*, vol. 10, no. 8, 2014.

[45] B. Ramanathan, H. M. Jindal, C. F. Le, R. Gudimella, A. Anwar, R. Razali, J. Poole-Johnson, R. Manikam, and S. D. Sekaran, "Next generation sequencing reveals the antibiotic resistant variants in the genome of pseudomonas aeruginosa," *PloS one*, vol. 12, no. 8, 2017.

[46] D. Botstein and D. Shortle, "Strategies and applications of in vitro mutagenesis," *Science*, vol. 229, no. 4719, pp. 1193–1201, 1985.

[47] H. J. Muller, "Artificial transmutation of the gene," *Science*, vol. 66, no. 1699, pp. 84–87, 1927.

[48] C. Auerbach, J. M. Robson, and J. Carr, "The chemical production of mutations," *Science*, vol. 105, no. 2723, pp. 243–247, 1947.

[49] A. J. Clark and G. J. Warren, "Conjugal transmission of plasmids," *Annual review of genetics*, vol. 13, no. 1, pp. 99–125, 1979.

[50] H. S. Seifert, E. Y. Chen, M. So, and F. Heffron, "Shuttle mutagenesis: a method of transposon mutagenesis for saccharomyces cerevisiae," *Proceedings of the National Academy of Sciences*, vol. 83, no. 3, pp. 735–739, 1986.

[51] N. Kleckner, J. Roth, and D. Botstein, "Genetic engineering in vivo using translocatable drug-resistance elements: new methods in bacterial genetics," *Journal of molecular biology*, vol. 116, no. 1, pp. 125–159, 1977.

[52] G. M. Blackburn, M. J. Gait, D. Loakes, D. M. Williams, M. Egli, A. Flavell, S. Allen, J. Fisher, S. I. Haq, J. W. Engels, *et al.*, *Nucleic acids in chemistry and biology.* Royal Society of Chemistry, 2006.

[53] D. Shortle, D. DiMaio, and D. Nathans, "Directed mutagenesis," *Annual review of genetics*, vol. 15, no. 1, pp. 265–294, 1981.

[54] B. Cunningham and J. Wells, "High-resolution epitope mapping of high-receptor interactions by alanine-scanning mutagenesis," *Science*, vol. 244, pp. 1081–1085, JUN 2 1989.

[55] D. M. Fowler and S. Fields, "Deep mutational scanning: a new style of protein science," *Nature Methods*, vol. 11, pp. 801–807, AUG 2014.

[56] D. Melamed, D. L. Young, C. E. Gamble, C. R. Miller, and S. Fields, "Deep mutational scanning of an rrm domain of the saccharomyces cerevisiae poly (a)-binding protein," *Rna*, vol. 19, no. 11, pp. 1537–1551, 2013.

[57] L. M. Starita, J. N. Pruneda, R. S. Lo, D. M. Fowler, H. J. Kim, J. B. Hiatt, J. Shendure, P. S. Brzovic, S. Fields, and R. E. Klevit, "Activity-enhancing mutations in an e3 ubiquitin ligase identified by high-throughput mutagenesis," *Proceedings of the National Academy of Sciences*, vol. 110, no. 14, pp. E1263–E1272, 2013.

[58] C. Li, W. Qian, C. J. Maclean, and J. Zhang, "The fitness landscape of a trna gene," *Science*, vol. 352, no. 6287, pp. 837–840, 2016.

[59] S. Webb, "Deep learning for biology," *Nature*, vol. 554, no. 7693, 2018.

[60] S. Russell and P. Norvig, "Ai a modern approach," *Learning*, vol. 2, no. 3, p. 4, 2005.

[61] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees. belmont, ca: Wadsworth," *International Group*, vol. 432, pp. 151–166, 1984.

[62] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[63] N.-L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, and P. C. Ng, "SIFT web server: predicting effects of amino acid substitutions on proteins," *Nucleic Acids Research*, vol. 40, pp. W452–W457, JUL 2012.

[64] M. Hecht, Y. Bromberg, and B. Rost, "Better prediction of functional effects for sequence variants," *BMC genomics*, vol. 16, no. 8, p. S1, 2015.

[65] P. Yue, Z. Li, and J. Moult, "Loss of protein structure stability as a major causative factor in monogenic disease," *Journal of Molecular Biology*, vol. 353, pp. 459–473, OCT 21 2005.

[66] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nature Methods*, vol. 7, pp. 248–249, APR 2010.

[67] T. Kawabata, M. Ota, and K. Nishikawa, "The protein mutant database," *Nucleic acids research*, vol. 27, no. 1, pp. 355–357, 1999.

[68] T. A. Hopf, J. B. Ingraham, F. J. Poelwijk, C. P. I. Scharfe, M. Springer, C. Sander, and D. S. Marks, "Mutation effects predicted from sequence co-variation," *Nature Biotechnology*, vol. 35, pp. 128–135, FEB 2017.

[69] A. J. Riesselman, J. B. Ingraham, and D. S. Marks, "Deep generative models of genetic variation capture the effects of mutations," *Nature Methods*, vol. 15, pp. 816+, OCT 2018.

[70] B. J. Livesey and J. A. Marsh, "Using deep mutational scanning data to benchmark computational phenotype predictors and identify pathogenic missense mutations," *BioRxiv*, p. 855957, 2019.

PART I

Correlated Mutations

Chapter 2

Statistical Characteristics of Amino Acid Covariance as Possible Descriptors of Viral Genomic Complexity

Abstract

At the sequence level it is hard to describe the complexity of viruses which allows them to challenge host immune system, some for a few weeks and others up to a complete compromise. Paradoxically, viral genomes are both complex and simple. Complex because amino acid mutation rates are very high, and yet viruses remain functional. Simple because they have barely around 10 types of proteins, so viral protein-protein interaction networks are not insightful. In this work we use fine-grained amino acid level information and their evolutionary characteristics obtained from large-scale genomic data to develop a statistical panel, towards the goal of developing quantitative descriptors for the biological complexity of viruses. Networks were constructed from pairwise covariation of amino acids and were statistically analyzed. Three differentiating factors arise: predominantly intra- vs inter-protein covariance relations, the nature of the node degree distribution and network density. Interestingly, the covariance relations were primarily intra-protein in avian influenza and inter-protein in HIV. The degree distributions showed two universality classes: a power-law with exponent -1 in HIV and avian-influenza, random behavior in human flu and dengue. The calculated covariance network density correlates well with the mortality strengths of viruses on the viral-Richter scale. These observations suggest the potential utility of the statistical metrics for describing the covariance patterns in viruses.

Our host-virus interaction analysis point to the possibility that host proteins which can interact with multiple viral proteins may be responsible for shaping the inter-protein covariance relations. With the available data, it appears that network density might be a surrogate for the virus Richter scale, however the hypothesis needs a re-examination when large scale complete genome data for more viruses becomes available.

2.1 Introduction

The genome size and complexities in different organisms vary widely. While bacteria have genes encoding several thousand types of proteins, most viruses have barely around ten types of proteins. This is true for viruses as benign as common flu to the lethal ones like ebola. As the number of base-pairs encoding these genes across the species varies from hundreds of millions to tens of thousands, the mutation rate which is the chance of making an error over a generation increases by many orders of magnitude.[1, 2] Despite this high rate of mutations or errors in the amino acids of viral proteins, many viruses remain functional and infect the hosts possibly because many deleterious mutations are compensated by other mutations. Continuously evolving viruses thus become much more unpredictable both for the immune system as well as the drugs developed against them. Characterizing the evolutionary behavior of viruses will thus be an important step towards understanding the complexity of viruses. Yet, to date there is no theoretical way of describing the complexity of viruses and their evolution.

One way of describing the systems-level complexity involved in healthy and diseased cells is by studying interaction networks. Biological networks can be formed out of transient molecular interactions such as in proteins interacting with other proteins.[3, 4] Metabolic[5] and gene regulatory networks[6] are other examples of functional cellular networks. Disease networks on the other hand try to connect genotypes with phenotypes.[7, 8] Protein-protein interaction networks have been used to describe the complexity of different systems from *E. coli* to humans.[9] Protein-protein interaction networks reveal several insights into the cellular functioning, such as the proteins in the network hubs which ubiquitously interact with other proteins, the evolutionary conservation of the networks across different species, and systems-level stability of the networks under removal of certain proteins.

Viruses have high rates of mutation, possibly arising out of their complex interactions with hundreds of human host proteins[10] during viral replication and pathogenesis.[11] Viral proteins evolve either to reduce certain interactions or to maintain them as the host proteins themselves undergo mutations.[12, 13] There is increasing number of studies that reveal these virus-host interactions. The focus of the present work is however, to statistically describe the viruses at the complete genome level, selecting a scale that is bigger than a single protein and smaller than the virus-host interactions. Since viruses have only around ten types of proteins, building interaction maps either at the protein level or at the domain level will have too little information to draw systems level inferences or to compare one virus with another. Since the uniqueness of viruses is their high mutation rates, fine-graining with a focus on amino acid interactions is statistically and biologically more meaningful.[14–17] Finer scale manifestations of protein-level interactomes[18] have been studied in the domain-level interactomes of *C. elegans*[19] as well as in the amino acid level interactions in viruses.[16, 17] The consequences of such studies span from the potential that it may be possible to define a systems-level metric for the viral complexity to identifying suitable strategies for drug discovery by highlighting the amino acid level interactions. In this work, we explore the former aspect on how viral or viral genomic complexity may be defined, a question that has not been asked so far to the best of our knowledge. Amino acid level covariance can arise either from structural constraints between proximal amino acids or because of functional constraints from amino acids at distal sites or other proteins or due to phylogeny.[20, 21] Several studies focused on building amino acid interaction networks, starting from the three dimensional structural data of proteins.[14, 15, 22] The utility of structure based methods is limited to availability of the structures, and to structurally proximal relations. Conversely, using amino acid co-evolutionary couplings from abundant homologous sequence data of multiple species,[23] bioinformatic approaches such as Statistical Coupling Analysis (SCA),[24] Direct Coupling Analysis[25] and GREMLIN[26] could predict hotspots of proteins, active sites of enzymes, *de novo* three dimensional structures,[27, 28] protein-protein contacts,[29] functionally related clusters of amino acids[30] and the vulnerability of viruses.[31] In this study, we use amino acid covariance networks from whole genome data to study the systems level characteristics of viruses. Earlier studies had explored and identified the genome-wide amino-acid co-variational couplings in various viruses.[17] The analysis was based on the smaller data sets available then, and the mechanism underlying the observed power-law, which is different from the ones in commonly studied complex networks, was not explored. In this work, we use large-scale complete genome data obtained from thousands of sequences of each virus to build amino acid covariance networks. We further use these network characteristics to probe the systems level complexity of the interaction networks, with possible implications for defining the biological complexity of viruses.

2.2 Results

2.2.1 Amino acid covariance networks

Degree of conservation is a statistical measure at individual amino acid level, and covariance is its extension to pairwise amino acid interactions. In this work we create a systems level extension of this pairwise covariance, the amino acid covariance network, which can represent the statistical nature of the variations in the complete viral genome across patients. Large scale genomic data of viruses was obtained from the NCBI servers (**Methods** section). With the current publicly available data, and our constraint that complete genome data from at least 1000 patients is available, only five viruses were chosen for analysis: HIV-1 subtype B (referred to as HIV), hepatitis subtype B (referred to as hepatitis), dengue, avian and human influenza subtype A (referred to as human influenza); however, the availability of such data is increasing. Multiple Sequence Alignment (MSA) of the complete genome data from all patients was performed. Using consensus sequence as a reference, the entire MSA was converted into a binary representation, 1 if the amino acid at a given position in a sequence is the same as that in the consensus sequence, 0 otherwise. Using the Statistical Coupling Analysis protocol,[24] weighted covariance matrix **C** that quantifies the relations among the different amino acids was created. The covariance matrix was further corrected for phylogeny effects by eliminating the component corresponding to the highest eigen value, as well by removing the modes with eigenvalues smaller than the eigenvalues of a random matrix (**Methods** section). Since the sequences for viruses which are from a cohort rather than across multiple species are closely related, the modes other than the first one also could have contribution from phylogeny and hence the covariance can have phylogenetic origin. The data on pairwise covariance was then converted into a network representation, where the amino acids form the nodes and the covariance relations form the edges or the connections between the nodes. The network representation allows visualization and analysis of the relations at a complete genome level, more intuitively than with covariance matrices, **C**. If in the covariance matrix any element C_{ij} relating amino acids i and j exceeds a threshold C, $|C_{ij}| > C^{th}$, then the covariance relation is considered to be significant and an edge i —j is created in the network. As it is demonstrated later, the threshold did not affect the broad statistical conclusions. The amino acid covariance networks for the viruses are shown in **Figure 2.1**.

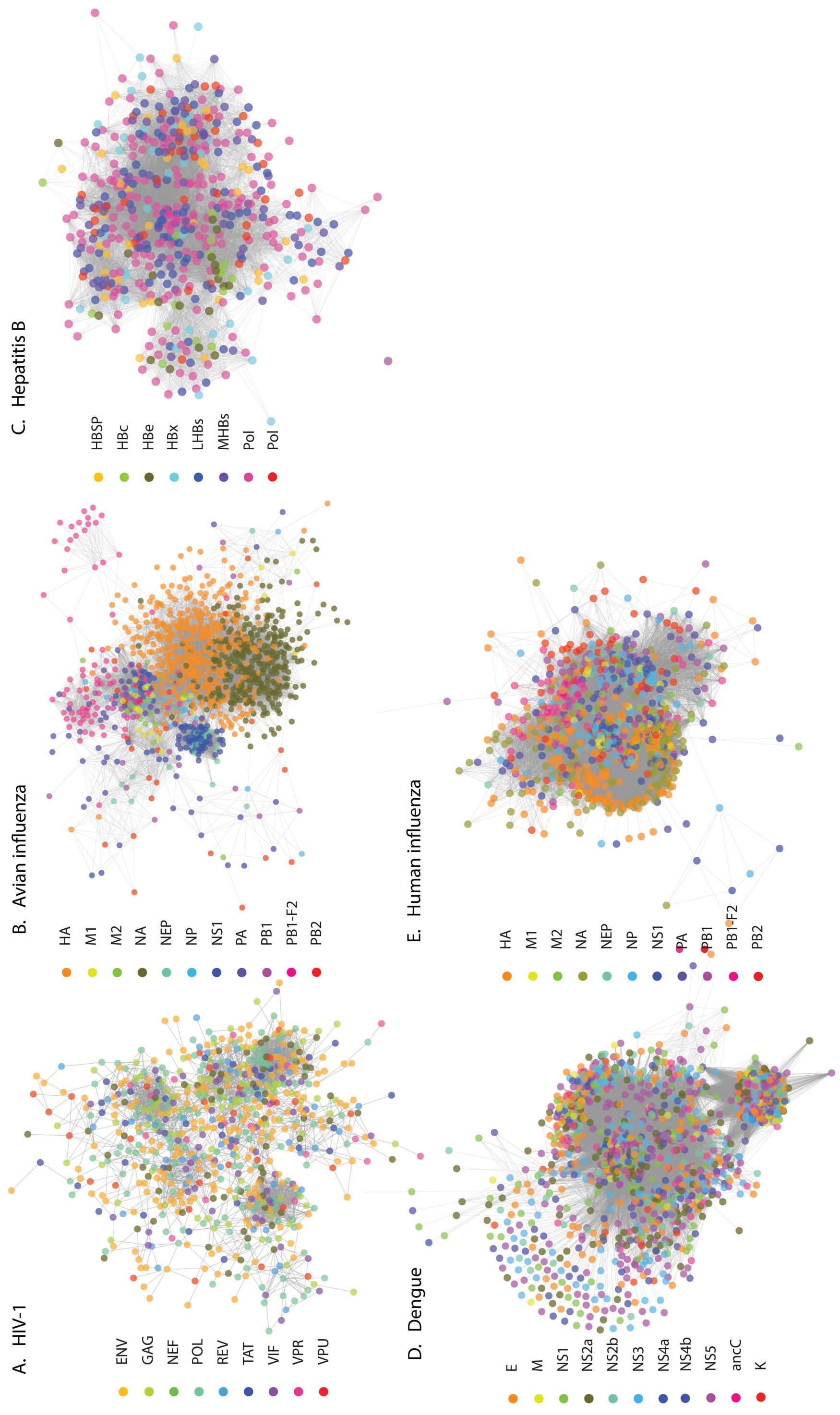

Figure 2.1: Covariance network from complete genome analysis of different viruses: (A) HIV, (B) avian influenza and (C) Hepatitis (D) Dengue (E) Human influenza. The networks are generated using covariance strength as a weight. The side bar indicates the different types of proteins found in these viruses, as well as the coloring notation used. The networks show three to four major clusters. While in HIV, each cluster has a mixed representation from all the proteins, avian influenza clusters are mainly from intraprotein covariance relations. Network representations were generated using Cytoscape. [32]

2.2.2 Intra-protein vs. inter-protein clusters

Using the complete genome data from different patients, the covariance networks for different viruses were constructed. We performed the Principal Component Analysis on the covariance matrix, rank ordered the eigenvalues and used Cattell's criterion[33] for noting the significant number of clusters. This criterion resulted in about 3 to 4 significant clusters for all the viruses. Clustering of nodes was also performed in Cytoscape software, using correlation as a weight (**Figure 2.1**) with the goal of observing patterns which are more general than those seen in pairwise relations and this analysis also resulted in 3 to 4 significant clusters. As can be seen, the amino-acid composition of each of the clusters in the viruses was noticeably different. In HIV the inter-protein covariance relations are much stronger. The same qualitative difference is quantitatively summarized by the number of connections within and between different proteins in Tables A.1 to A.5 of Appendix A. The summary of the fine-grained inter- versus intra-protein covariance relation strengths in each of the clusters is visualized as chord-diagrams and the compositions of the clusters from different proteins are represented in Figures A.1 to A.5 and A.6 respectively (Appendix A). Seen at the protein-protein interaction level, clearly there are interactions between any pair of proteins, however, the differences in the numbers of interactions or the overall strengths come from basing the analysis at an amino acid level. For dengue, hepatitis and human influenza also the clusters have residues from multiple proteins.

2.2.3 Node degree distribution

One advantage of transforming the covariance matrix into a network is that several systems-level statistical analyses can be performed. The complexity of the networks is analyzed by studying its node-degree distribution, $n(k)$ - the number of times a node with a certain number of edges k appears in the network.[34] Two commonly observed universality classes in these distributions - power-law and Poissonian, suggest a systematic or random underlying basis,[34] and these occur in the amino acid degree distributions as well. In HIV, power-law $n(k) \sim k^{-\gamma}$, $\gamma \sim 1$, was significantly observed, while dengue and human influenza show random distribution (**Figure 2.2**). Hepatitis and avian influenza on the other hand showed a mixed behavior including both powerlaw and random behaviors (**Figure 2.2**). We further analyzed the role of the threshold by varying C^{th} in the analysis of hepatitis. As shown in **Figure 2.3**, as the C^{th} was increased from 0.50 to 2.0, the powerlaw component becomes more pronounced (similar data for other viruses is shown in Figures A.7 to A.10 of Appendix A. The data shows a clear separation of network connections arising from two different origins, an organized

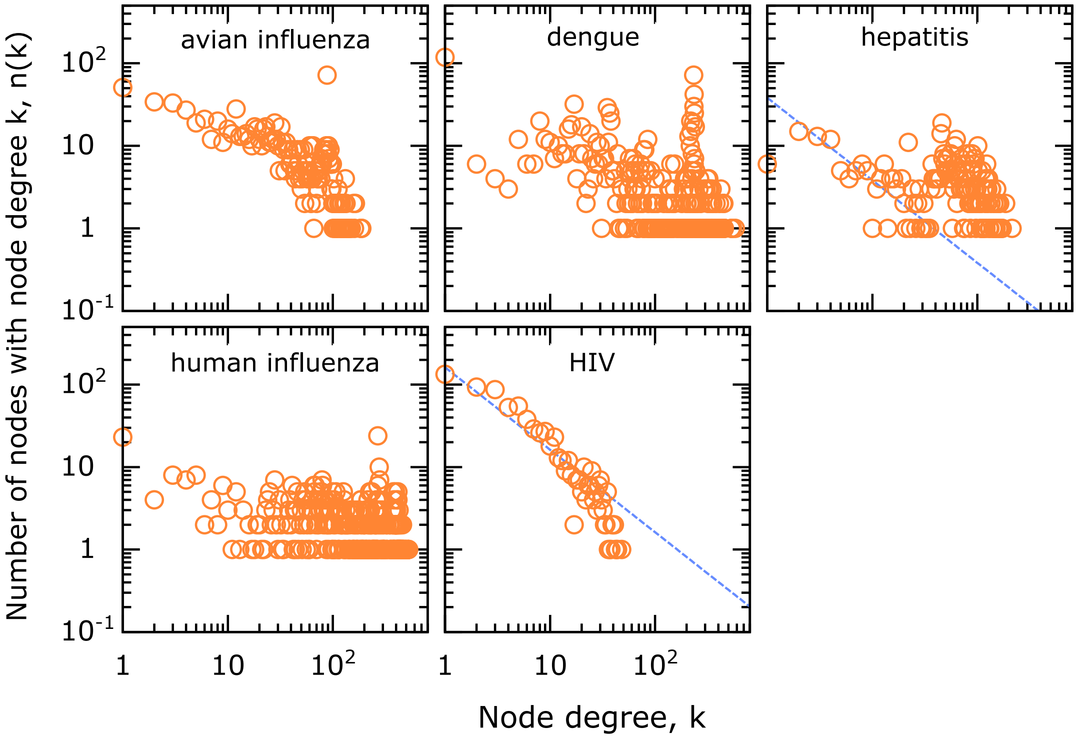

Figure 2.2: Node degree distribution from the complete genome data in different viruses showing a range of behavior from a pure power-law (HIV) to a pure-random network behavior (dengue and human influenza). The dashed line in the panels for hepatitis and HIV is shown for reference and corresponds to power-law with exponent -1. A cutoff $C^{th} = 0.7$ was used as a threshold for establishing network edge connections. The effect of changing the cutoff is discussed separately.

network of covariance above a certain threshold and random network connections at lower thresholds of covariance. Within this powerlaw regime a further change in cutoff did not result in a change in the exponent significantly. We also performed another simple phylogenetic check by comparing the analysis on dengue serotype 1 (Figure A.11 of Appendix A), with that on the combined data from all dengue serotypes. While this analysis does not prove that the phylogenetic effects were negligible, it does suggest that even strong phylogenetic corrections such as performing the analysis only on one subtype did not change the conclusions.

The analysis presented so far is the statistical description of data collected from patients and is averaged over all the years of sample collection. In order to study the temporal evolution patterns, we performed time analysis on the data set which is most abundant, human influenza (subtype A). We divided the complete genome data from human influenza into periods where the number of sequences is similar ($\sim$ 2000 complete genomes each). A node-distribution analysis shows that over this period, there is no significant change in the covariance complexity of viruses (Figure A.12 of Appendix A).

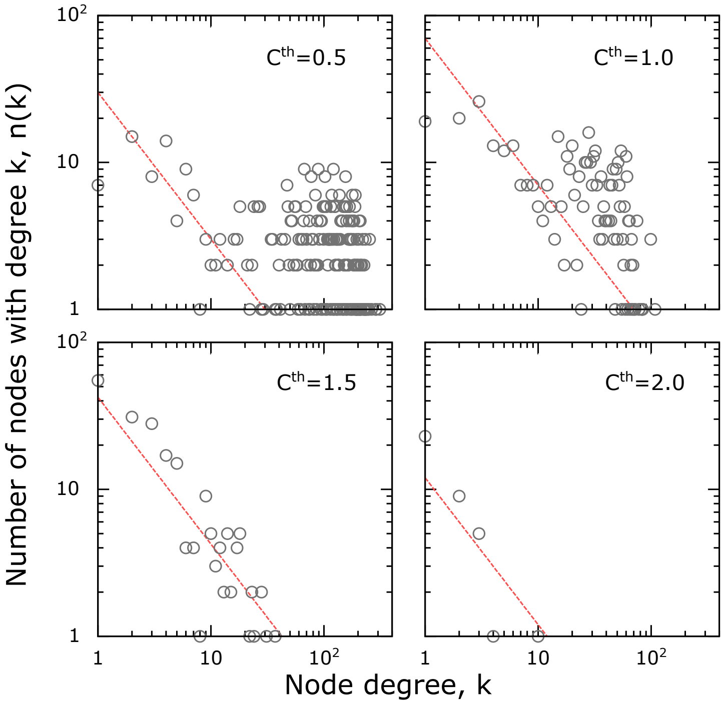

Figure 2.3: Node degree distribution sensitivity was studied in hepatitis network by changing the cut-off value used for defining edge connectivity between the nodes. At a very low cut-off there is a mixed behaviour in the node degree distribution, with both power-law as well as a random component. As the cut-off is increased, the random component is selectively removed, while preserving the power-law component. This suggests a clear separation of network connections from random and systematic origins. By choosing a threshold value, one can filter and study just the systematic component. The dashed line corresponding to power-law with exponent -1 is drawn for reference.

2.2.4 Network density

Network density is the fraction of the edges (connections) between the nodes in a network relative to the total number of edges possible between the nodes from purely combinatorial considerations that edges can be formed between any pair of nodes. The densities were calculated using Cytoscape software, and they range from sparse to dense networks. These parameters are related to the qualitative nature of the node-degree distribution, as the sparse networks tend to be scale-free, while dense networks are more likely to be random networks. In fact, network density parameter quantified the transition from different degrees of randomness to systematic connections which result in power-laws, and we wanted to compare this with a known metric of biological complexity. The only scale that we are aware of, that makes a direct comparison between the impact of different viruses is the Virus Richter scale,[35] which ranks viruses according to the logarithm of the mortality they cause. The network density for each virus was calculated by choosing the threshold which was the cusp of the transition between random to powerlaw behaviors.

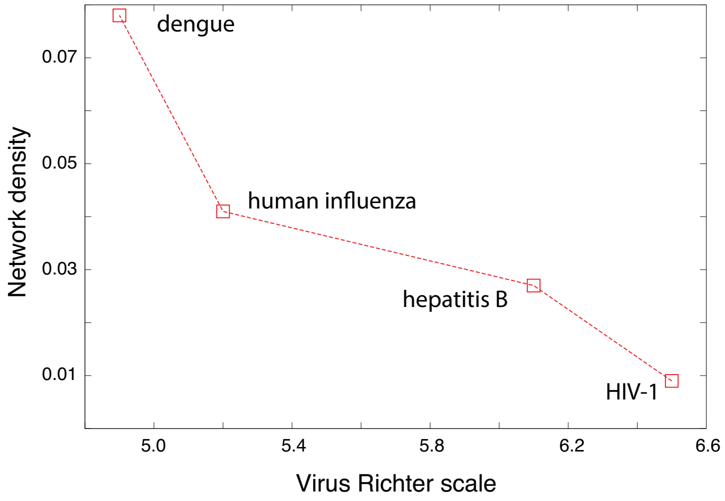

Figure 2.4: The relation between the complexity of the virus, as described by Virus Richter scale, and its network characteristic - density (□)

The network density from our calculations was plotted against the virus Richter scale in **Figure 2.4**, and the two are anti-correlated with a Pearson correlation -0.929 ($p \approx 0.07$).

2.2.5 Robustness of networks

In typical network analyses, pairwise relations are used for constructing the network, and the systems-level statistical properties are interpreted from it. As such it is important to see the effect of the removal of a few nodes and the edges connected to them.[36] The change in the system level properties such as network diameter on the removal of a few nodes has been interpreted as the sensitivity of the network to a random or targeted attack.[36] We checked for the robustness of amino acid covariance network by removing different fractions of nodes and all the edges connecting to them, the spirit being that the critical amino acids or groups of them can be a potential drug target. The nodes to be removed were chosen according to two strategies: (a) randomly or (b) by picking those with the highest degree, to simulate a random error or a targeted attack, **Figure 2.5** shows how the effective diameter - a metric of network connectivity - is affected by the targeted or random removal. Targeted removal has the highest effect on HIV followed by avian influenza. For these two virus covariance data sets, the difference between targeted and random removal of nodes is significant, compared to all other viruses. The disruption of the network in the case of HIV, with the removal of a small fraction of the nodes, suggests that very few nodes act as hubs and moderate most of the interactions in the network. The overall characteristics of robustness may be intuitively expected from the the power-law distribution of nodes.

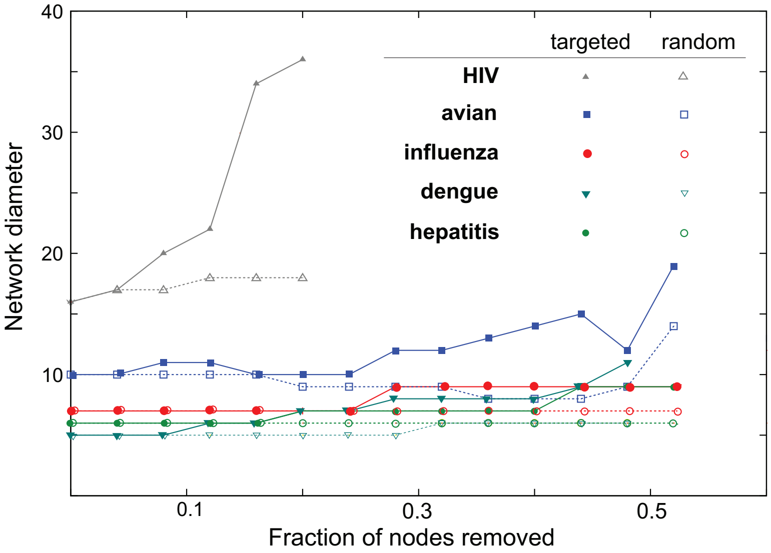

Figure 2.5: The robustness of the networks is studied by calculating the change in the network diameter in response to targeted and random removal of nodes.[36] HIV and avian-influenza data show a significant difference between targeted and random removal, the latter being much lower, suggesting that these networks can be destabilized more by a targeted attack. After removing a very high fraction of nodes, networks breaks down into smaller disconnected clusters, resulting in a decreased diameter, and this part of the data where the network is heavily destabilized is not shown in this graph.

2.2.6 Powerlaw exponent

The powerlaw exponents, $\gamma \sim 1$, observed in our study is different from the usually observed power-laws with $\gamma \sim 2-3$ for which there are several mechanistic explanations including influencer models.[34] In our analysis the exponent was also robust to halving the data sets, and needed an alternative interpretation relevant for covariance. Considering amino acid conservation (ϕ) as a surrogate for their fitness, we developed a fitness based model.[37] The model uses two distributions derived from the whole genome data: (a) the distribution of the conservation among the amino acids, $p(\phi)$ (Figure A.13 of Appendix A) (b) the covariance fitness potential of the node $\eta(\phi)$ corresponding to a given conservation of the amino acids. The latter can be modeled as a gaussian distribution, with minimal covariance fitness for amino acids with very high and very low conservation, a peak in between at ϕ_m and standard deviation σ. Considering a pair of amino acid nodes i and j, and two random numbers r_1 and r_2 drawn from a uniform distribution, edge i —j is created in our model if $r_1 * r_2 \leq \eta(\phi_i) * \eta(\phi_j)$. This algorithm generates a node-degree distribution with $\gamma \sim 1$ (Figure A.14 of Appendix A). The model explains power-law with exponent $\gamma \sim 1$, random distribution, and a transition to the powerlaw, as seen in hepatitis (**Figure 2.3**). For example, for HIV, the conclusion is relatively invariant for a gaussian with $\phi_m = 0.6 - 0.7$ and $\sigma = 0.02 - 0.07$. As the parameters go out of this range, node degree distribution eventually transforms to a random network model.

2.2.7 Correlation with host protein interactions

We examined the possible relation of covariance couplings to host-virus interactions, with the interactomes from dengue, human influenza and HIV-1. Two different comparisons were made: (i) the number of common host proteins between a pair of proteins and the total number of inter-protein covariance couplings for this pair (Figure A.15 of Appendix A) (ii) the importance of a viral protein in the combined virus-host interactome, quantified by the eigenvector centrality, and the number of total covariance couplings a protein has (Figure A.16 in Appendix A). Other centrality measures were also analyzed, but there was no difference in the conclusions. The two different comparisons showed correlation between the number of covariance couplings and the strength of interactions in the interactome. The same pattern could not be seen in the interactome data we used for the other viruses. With the data available, the viral interprotein interactions were classified as direct, indirect mediated by host proteins, and non-existent (Figure A.17 of Appendix A), but no clear inference could be drawn. We performed a complementary analysis by counting the number of viral proteins that each host protein interacts with. The analysis represented in **Figure 2.6** shows that the viral proteins are clustered closely in dengue and influenza interactomes because many of the host proteins interact with more than one viral protein, making the couplings stronger.

2.3 Discussion

Amino acid mutations are robustly networked: Mutations occur very frequently among viral proteins. Yet among these variations occurring at different sites, in different viral proteins, there are interdependencies. Most co-evolution or covariance based studies focused on bacterial proteins, and very few on viral proteins. Some examples are of intraprotein co-evolutionary interactions in the GAG polyprotein of HIV subtype B,[31] with the goal of identifying collectively-coordinated functional units within these proteins, as well as the co-variation networks in genome wide virus data.[17] While interesting questions on genome-wide relationships among different viruses had been raised in that work, in a similar spirit as the present work, the analyses were based on less than hundred sequences. Several issues remained unclear - the sensitivity of these analyses to larger data sets, to a different choice of the definition of covariance, the origins of power-law and possible connections to the biological complexity of viruses. These are the questions we explored in this work. Even with the choice of larger data sets, the covariance relations remained.

Almost all the networks are robust to the random or targeted removal of about 10%

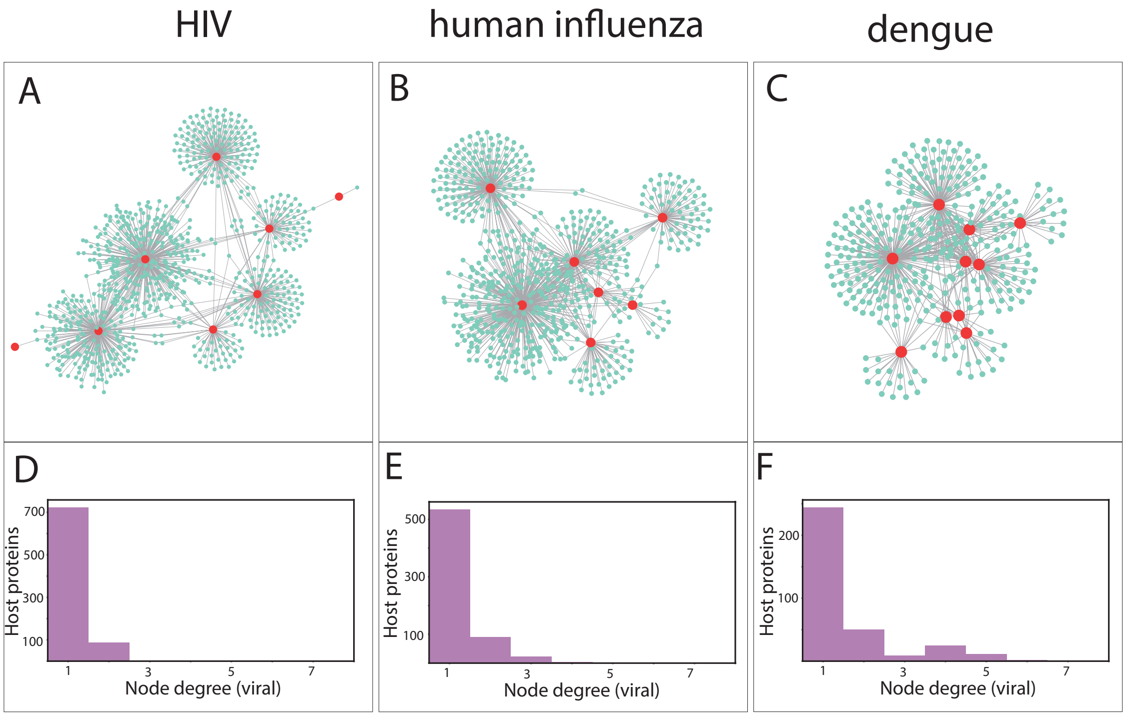

Figure 2.6: Analysis of host-virus interactions was performed using the interactomes of human protein with HIV, dengue and human influenza (details in **Methods** section). The networks of interactions are shown in (A), (B), (C) with the host and viral proteins represented in cyan and red colors respectively. The number of viral proteins any given human protein is interacting with is denoted as the node-degree (viral) and the number of such proteins in the interactome are indicated on the y-axis in (D), (E), (F). HIV has the highest number of proteins interacting with a single viral protein and dengue has the largest number of host proteins interacting with more than one viral protein.

of nodes and they start showing differential behavior beyond this (**Figure 2.5**). The differences in network characteristics relative to a random or a targeted node removal (**Figure 2.5**), combined with one of the interpretations in the network theory,[36] leads us to a possible hypothesis. Scale free (powerlaw) networks were originally speculated to be stable against any attack, and only later[36] it was learnt that while this may be true under a random attack, these networks are vulnerable to a targeted attack. A possible inference, specifically for the viral complexity, is that the viruses with powerlaw covariance networks may be vulnerable to an attack on a group of their amino acids in a targeted manner. This inference is conceptual in nature, suggesting that there might be a better way to design drugs targeting even these otherwise complicated viruses. The drug targets may be chosen from either the same protein or from multiple proteins, as suggested by the strength of intra- or inter-protein interaction networks. However, the practical choices of drug targets, chosen from the networks, and the possible consequences are out of the scope of the present work.

Networks are statistically significant: Most co-evolution studies focused on using homologous sequences of bacterial proteins originating from different species for their analysis, and required the number of sequences[30, 38] to be anywhere from 100 to 1000. Without highlighting the mathematical details, using a metric of distance between sequences, a concept of effective sequences was introduced[30, 38] to discount the sequences that are close to one another within the same cluster of sequences and but are further apart from the other clusters. The sequences which are with an identity better than 0.8 were effectively considered to be the same sequence, thus weighing down the total number of sequences. The present analysis is different from the commonly used co-evolution studies in several ways: (i) Sequences are from within the subspace of the same virus, representing polymorphisms, rather than from the hypothetical sample set from all viruses or all proteins. Thus the sequence identities are high and a cutoff of 0.8 was not relevant (ii) Further, weighting of the sequences was not used in our covariance network generation (iii) By choosing increasing homology cut-offs over 0.9 (Figure A.18 of Appendix A), which are still relevant for the virus polymorphisms, the number of effective sequences increases over 100. We thus believe that the size of the sequence data sets used was sufficient, although it might appear to be insufficient based on the standard definitions of number of effective sequences. Further, to eliminate the possibility that the observed patterns in the node degree distribution are an artefact because of the higher number of effective sequences of HIV and avian influenza, the covariance analysis was repeated using randomly selected 200 sequences from the alignment. Even with this significantly reduced number of sequences, the statistical nature of the couplings did not change for HIV and avian influenza. The characteristics of the distribution remained the same for all the viruses as shown in Figure A.19 of Appendix A.

We further verified the statistical significance and reliability by (i) halving the number of sequences, which did not change the conclusions (ii) evaluating the p-value of the connections, which for all the connections turned out to be < 0.01. We also repeated the analyses separately on the raw covariance matrix. Although the number of connections drastically increased compared to that when the cleaned matrix was used, the statistical characteristics such as powerlaw dependence and the anticorrelation with the virus Richter scale did not change (data not shown). Understanding that several eigenvalues, not just the first one could be contributing to phylogenetic effects,[21] we repeated the calculations by removing the contributions from top 5 and 10 eigenvalues until the networks had very few connections. The results shown in Figures A.20 and A.21 of Appendix A suggest that the qualitative patterns of powerlaw and random network did not change. We thus believe that the covariance connections observed in our analysis were statistically significant.

The analysis was repeated using an alternative method, MaxSubTree,[39] for identifying the covariance relations. The two objectives of this investigation were to use a method that is suitable for finding co-evolving or covarying residues from sequences with variable divergences[39] and also to show that the topology of the covariance network is not sensitive to the choice of our method. As dengue virus had the least diverged sequences, the analysis was performed using the publicly available code for MaxSubTree.[39] We observed random topology for the covariance network generated using this method also (Figure A.22 of Appendix A).

Covariance is related to conservation: The general pattern in node-degree distribution was that some networks are scale-free with powerlaw distribution and others are random networks. In fact, it was seen that two different classes of covariance, scale-free and random component, were simultaneously present and the scale-free component became significant at higher thresholds (C^{th}) for some viruses. While the formation of random network connections at lower thresholds may be expected, having powerlaw distributed patterns at higher thresholds is non-trivial and we discuss further about a possible explanation below.

While the covariance networks can be statistically described using scale-free or random node-degree distributions, insights into the covariance come from the observed exponent, $\gamma \sim 1$, in the scale-free distribution. Random networks (Erdos-Renyi model), small world networks (Watts-Strogatz model [40]) and self-similar networks (Barabasi-Albert model[41, 42]) arising in diverse contexts such as WWW, protein-protein interactions, citation networks, etc have been well studied. The powerlaw with $\gamma \sim 1$ observed in the covariance network is different from the typical powerlaws γ varying from 2 to 3 and is closer to the behavior in co-authorship networks. Some of the mechanisms that explain the observed phenomena are preferential attachment model[34] where newer edges are added to a node depending on its current degree, or based on its pre-defined fitness or a potential for a degree. Unlike a citation network, there is no reason to believe that the covariance network evolves with a continuous increase in the number of nodes and edges. In the model presented in this work, powerlaw with exponent $\gamma \sim 1$ was derived assuming that the covariance between a given pair of amino acids depends simultaneously on the conservation of both these amino acids under consideration. The model captures the observed powerlaw with the minimal assumption that the covariation of a pair of amino acids is related simultaneously to their conservations, which seems plausible.

Comparative mortality from viruses: An important question to pursue is about why the human immune system finds it easy to fight certain infections and not others. On the surface, defining the complexity of the viral infections seems plausible because the viral genome is relatively simple, and is about 1000 times smaller than the bacterial genome.

An attempt to define and quantify the complexity of the viral genome seems relevant and timely, especially since the genomic data is becoming readily available. However, it is difficult to describe complexity, and even more to quantify it with one single measure. The lack of a simple and precise metric for complexity is a challenge both in biology as well as from theoretical calculations. For biological complexity of viruses, here we use the strength on virus Richter scale[35] as a surrogate measure. Virus Richter scale indicates mortality from viruses, which implicitly includes several factors from how fast the virus mutates to how poor the public health provisions are. We use virus Richter scale as, to the best of our knowledge, there is no other metric comparing the strengths of viruses or difficulty of developing vaccines against them. **Figure 2.4** shows a plot between the virus strength and the network characteristic - network density. Richter scale data for avian influenza was not available and hence was not included in this analysis. The observed anti-correlation between the network density which is a network metric and the biological metric is obtained from just four viruses ($p = 0.07$), and needs to be re-evaluated when further data becomes available. However, it raises the possibility that the complexity of the biology and the pathogenicity of the virus may be reflected in the amino acid covariance networks.

Node-degree distribution of the covariance networks, depending on the virus, was demonstrated to assume qualitative patterns ranging from predominantly powerlaw to a predominantly random network distribution. It was also clear from the results that the random component quantitatively has a higher contribution to the node degree. Thus the higher values of the network density in **Figure 2.4** reflect higher contributions from the random components, and the reducing network density describes the transition from primarily random network to one with a powerlaw. The former type of network was seen more sensitive to random attacks (**Figure 2.5**), which offers a plausible thread of logic for why with the continuously decreasing network density, decreases the randomly networked connections making the overall network of interactions resilient to random attacks on them.

Classifying the complexity of viral genomes: One might also have a similar feeling for which viruses are complex: either by examining the phylogenetic trees of the evolved sequences (Figures A.23 and A.24 of Appendix A) or even simply by knowing the time since when they infected the hosts: Influenza and hepatitis infections go back to thousands of years, the youngest among dengue serotype strains is about 200 years old and HIV and avian influenza are relatively younger with less than hundred years of exposure to their human hosts. Other works in the literature[43] have clustered viruses based on the shape of phylogenetic tree and found HIV and hepatitis C virus clustered together while dengue and human influenza A appeared in another cluster along with many other viruses. Thus introducing the network based analysis may at first seem redundant. However, the present work aims at developing several comparative measures between different viruses. Three different metrics were used, two of them qualitative: (1) are the amino acid covariance relations primarily intra-protein or mixed? (2) Is the node degree distribution scale-free or does it form a random network and (3) a quantitative measure of the network density. The complete genomic data from the five different viruses can be classified according to these metrics, and an anti-correlation with the viral Richter scale and the network density could also be observed. The standard deviation of the pairwise identities in the sequence data was also found to be negatively correlated with virus richter scale (Figure A.25 of Appendix A). The work thus raises questions on whether these statistical parameters can be used for describing the whole genome level viral evolution, distinguishing the viruses and the possibility to correlate these statistical metrics to the complexity of the viruses. When more data on viruses becomes available, it remains to be seen whether these three metrics are sufficient to classify the genomic complexity of viruses. The work also raises the possibility that by a suitable choice of target amino acids from the networks of covariance, it may be possible to destabilize the networks of even the complex viruses, with possible implications for drug discovery.

Host-virus interactions may be responsible for interprotein interactions: Interaction with host machinery and adaptation is an inevitable part of the virus infection cycle.[11] The multitude of coevolutionary relations among viral proteins could be arising out of direct interactions among themselves as well as because of the common interaction partners in the host. These interaction networks involve hundreds of human proteins[10] and viral proteins adapt with mutations in these host proteins.[12, 13] We investigated the possible correlation of number of common interaction partners and the number of covariance connections for protein pairs for HIV, human influenza and dengue (**Methods** section) and is shown in Figure A.15 of Appendix A. The positive correlation between the strength of covariance couplings and the number of common interacting host proteins in dengue virus was partly reassuring about the utility of covariance method, although the no pattern could be seen in the other two virus-host interactions.

By studying the host-virus protein interactome, we could observe (**Figure 2.6**) that dengue and human influenza have host proteins which interact with more than one viral protein. The evolution of viral proteins under the selection pressure from the host-virus protein-protein interaction may thus lead to a higher level of randomization in the interactions compared to HIV, where a very large number of host proteins interacted mostly with a single viral protein.

2.4 Conclusions

By using a network representation of amino acid covariance we had seen three different characteristics in the large scale complete genome data -a differentiable clustering with significant intra-protein or inter-protein couplings, the node degrees which have a structured power-law or random origins and the network density parameter. When genomic data from more viruses becomes available, it will be interesting to see if these three different measures of statistical complexity of genomes can be used to classify viruses into different categories, with a possible mapping to their biological or pathogenic complexity. Further it will be interesting to see if the inter-protein or intra-protein couplings are related to the host adaptation (HIV) or the host being a neutral carrier (avian influenza) and how such patterns evolve with time as the viruses adapt from being pandemics to epidemics.

2.5 Methods

Sequence selection: The complete genome data was curated from publicly available databases. With the two constraints that the complete genome data has to be available, and the number of sequences have to be more than 1000, we identified five viruses from the NCBI servers (https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?opt=virus&taxid=10239&host=human#). The individual protein data from different samples are available at the NCBI servers. However it was convenient to work with sources where the data curated by patient identity. The complete genome datasets available in the protein format were downloaded from different sources: HIV (http://www.hiv.lanl.gov), dengue (https://www.viprbrc.org/brc/home.spg?decorator=flavi_dengue), hepatitis (https://hbvdb.ibcp.fr/HBVdb/HBVdbDataset?seqtype=2), human and avian influenza (http://platform.gisaid.org/). Any sequence where information about all the proteins was not available was deleted from the analysis.

Multiple Sequence Alignment: Multiple sequence alignment of the curated sequences was performed using Clustal-Omega. Sequences which had a gap frequency more than 20% were excluded from the analysis.

Consensus sequence: The consensus sequence for each virus was generated using the most occurring amino acid at every given position. Using this sequence as a reference, the entire complete genome dataset was converted into a binary format: 1 if the amino acid in a given sequence matches amino acid at the corresponding position in the consensus sequence. This binarization or creating boolean strings is similar to the method used in Statistical Coupling Analysis,[30] which identified several functional relations among different amino acids.

Covariance networks: The chance of covariation C_{ij} between a pair of amino acids i and j is calculated by averaging the columns i and j of the boolean sequences using either an unweighted or weighted protocol following the Statistical Coupling Analysis protocol.[30] Unweighted and normalized covariance is defined as:
$C_{ij}^{unweighted} = (\langle x_i x_j \rangle_s - \langle x_i \rangle_s \langle x_j \rangle_s) \,/\, \left(\sqrt{\langle x_i^2 \rangle_s - \langle x_i \rangle_s^2} \sqrt{\langle x_j^2 \rangle_s - \langle x_j \rangle_s^2} \right)$, where x_i is the i^{th} column in the boolean sequence and $\langle\rangle_s$ denotes the average over sequences. Weighted covariance is defined as
$C_{ij}^{weighted} = \phi_i \phi_j \ |\langle x_i x_j \rangle_s - \langle x_i \rangle_s \langle x_j \rangle_s|$, where $\phi_i = ln\left((\langle x_i \rangle_s (1 - q^{a_i})) \,/\, (q^{a_i} (1 - \langle x_i^s \rangle_s)) \right)$, and q^{a_i} is the probability with which the amino acid a_i at position i in the consensus sequence occurs among all proteins. In the present work we use $C_{ij}^{weighted}$ and an undirected network link $i - j$ is created if C_{ij} exceeds a chosen cutoff c. The sensitivity of the analysis to c is discussed in the article.

Spectral cleaning: Since the correlation matrix C is symmetric, its eigenvalues are real and the eigenvectors can be used for spectral decomposition as: $C = \sum_k \lambda_k |k\rangle\langle k|$. The component corresponding to the highest eigenvalue of the correlation matrix is the contribution from phylogeny and is removed. Also the contribution from all the components having eigenvalues smaller than the second highest eigenvalue of the correlation matrix of randomized alignment is removed. So the cleaned correlation matrix is: $C_{cleaned} = \sum_{k=2}^{r} \lambda_k |k\rangle\langle k|$ where $\lambda_2 > \lambda_3 > ...\lambda_r > \lambda_{Ran}$. λ_{Ran} is the limiting value of the eigenvalue from the continuum of eigenvalues expected for the random matrix.

Network parameters Most network analyses, such as obtaining node degree distribution, clustering, network density were performed using Cytoscape[32]. Network diameter was calculated using NetworkX module of Python[44].

Clustering : We have used prefuse force directed layout with covariance as the edge weight for visualizing the covariance networks. In this layout, communities appear as groups of nodes,[45] hence it helps in identifying the community structures in networks.

Cattell's criterion: The eigen values of the correlation matrix was sorted in the descending order and plotted. The number of clusters is determined as the number of eigen values preceding the sharp change in the eigen values[33].

Robustness of network: 1) Error or Random removal : Nodes were selected randomly and removed. All the edges connecting to them were also removed. 2) Trageted attack : The nodes were sorted according to degree and the nodes with higher degree were removed first.

Number of effective sequences: Number of effective sequences was calculated as $N(I) = \sum_{k=1}^{n} 1/N_k$ where N_k is the number of sequences having identity $> I$ with the k^{th} sequence and n is the total number of sequences in the alignment. It was calculated before binarizing the alignment.

Virus-Host interactions: For virus-host interactions we found the most comprehensive data for: HIV, human influenza and dengue and we present the analyses for the same. The protein-protein interactions in the virus-host system was downloaded from virus mentha[46] (https://virusmentha.uniroma2.it/) for human influenza and HIV. For dengue virus the interactions with human host were obtained from DenvInt[47] (https://denvint.000webhostapp.com/index.html) as it had more records. Human protein interactome was obtained from mentha[48] (http://mentha.uniroma2.it/). The centrality measures for the combined protein interaction network of virus and host was calculated using the Networkx module of Python.

Chord diagrams: Chord diagrams showing protein-protein interaction strengths were prepared using the online tool Circos (http://circos.ca/)[49].

Phylogenetic tree: Rooted binary phylogenetic trees for all viruses were created in Matlab2017 using the Bioinformatics Toolbox. The seqlinkage function was used with the Jukes-Cantor pairwise distances between sequences.

MaxSubTree analysis: MaxSubTree[39] being a combinatorial approach can identify co-evolving amino acids from sequence alignments having variable divergence. The program is available at http://www.ihes.fr/,carbone/data7/MaxSubTree.tgz.

Data Availability: The datasets and the codes used for the analyses in the present study are available at https://doi.org/10.17605/OSF.IO/S3VUB

Bibliography

[1] J. W. Drake, “A constant rate of spontaneous mutation in DNA-based microbes ,” *Proc. Natl. Acad. Sci. USA,*, vol. 88, pp. 7160–7164, AUG 1991.

[2] R. Sanjuan and P. Domingo-Calap, “Mechanisms of viral mutation,” *Cellular and Molecular Life Sciences*, vol. 73, pp. 4433–4448, DEC 2016.

[3] J. Wang, S. Rao, J. Chu, X. Shen, D. N. Levasseur, T. W. Theunissen, and S. H. Orkin, “A protein interaction network for pluripotency of embryonic stem cells,” *Nature*, vol. 444, no. 7117, pp. 364–368, 2006.

[4] G. Kar, A. Gursoy, and O. Keskin, “Human cancer protein-protein interaction network: a structural perspective,” *PLoS Comput Biol*, vol. 5, no. 12, p. e1000601, 2009.

[5] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, “Network motifs in the transcriptional regulation network of escherichia coli,” *Nat. Genet.*, vol. 31, no. 1, pp. 64–68, 2002.

[6] G. Karlebach and R. Shamir, “Modelling and analysis of gene regulatory networks,” *Nature Reviews Molecular Cell Biology*, vol. 9, no. 10, pp. 770–780, 2008.

[7] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabasi, “The human disease network,” *Proc. Natl. Acad. Sci. USA*, vol. 104, pp. 8685–8690, MAY 22 2007.

[8] A.-L. Barabasi, N. Gulbahce, and J. Loscalzo, “Network medicine: a network-based approach to human disease,” *Nat. Rev. Genet.*, vol. 12, pp. 56–68, JAN 2011.

[9] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. Wanker, “A human protein-protein interaction network: A resource for annotating the proteome,” *Cell*, vol. 122, pp. 957–968, SEP 23 2005.

[10] B. De Chassey, L. Meyniel-Schicklin, A. Aublin-Gex, V. Navratil, T. Chantier, P. Andre, and V. Lotteau, “Structure homology and interaction redundancy for discovering virus–host protein interactions,” *EMBO reports*, vol. 14, no. 10, pp. 938–944, 2013.

[11] S. Khadka, A. D. Vangeloff, C. Zhang, P. Siddavatam, N. S. Heaton, L. Wang, R. Sengupta, S. Sahasrabudhe, G. Randall, M. Gribskov, *et al.*, “A physical interaction network of dengue virus and human proteins,” *Molecular & Cellular Proteomics*, vol. 10, no. 12, pp. M111–012187, 2011.

[12] M. D. Daugherty and H. S. Malik, “Rules of engagement: molecular insights from host-virus arms races,” *Annual review of genetics*, vol. 46, pp. 677–700, 2012.

[13] A. F. Brito and J. W. Pinney, “Protein–protein interactions in virus–host systems,” *Frontiers in microbiology*, vol. 8, p. 1557, 2017.

[14] G. Amitai, A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger, and S. Pietrokovski, “Network analysis of protein structures identifies functional residues,” *Journal Molecular Biology*, vol. 344, no. 4, pp. 1135–1146, 2004.

[15] K. Brinda and S. Vishveshwara, “A network representation of protein structures: implications for protein stability,” *Biophysical Journal*, vol. 89, no. 6, pp. 4159–4170, 2005.

[16] R. Aurora, M. J. Donlin, N. A. Cannon, J. E. Tavis, and V.-C. S. Grp, “Genome-wide hepatitis C virus amino acid covariance networks can predict response to antiviral therapy in humans,” *J. Clin. Invest.*, vol. 119, pp. 225–236, JAN 2009.

[17] M. J. Donlin, B. Szeto, D. W. Gohara, R. Aurora, and J. E. Tavis, “Genome-Wide Networks of Amino Acid Covariances Are Common among Viruses,” *J. Virol.*, vol. 86, pp. 3050–3063, MAR 2012.

[18] S. Li, C. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P. Vidalain, J. Han, A. Chesneau, T. Hao, D. Goldberg, N. Li, M. Martinez, J. Rual, P. Lamesch, L. Xu, M. Tewari, S. Wong, L. Zhang, G. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. Gabel, A. Elewa, B. Baumgartner, D. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S. Mango, W. Saxton, S. Strome, S. van den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. Gunsalus, J. Harper, M. Cusick, F. Roth, D. Hill, and M. Vidal, “A map of the interactome network of the metazoan C-elegans,” *Science*, vol. 303, pp. 540–543, JAN 23 2004.

[19] M. Boxem, Z. Maliga, N. Klitgord, N. Li, I. Lemmens, M. Mana, L. de Lichtervelde, J. D. Mul, D. van de Peut, M. Devos, N. Simonis, M. A. Yildirim, M. Cokol, H.-L. Kao, A.-S. de Smet, H. Wang, A.-L. Schlaitz, T. Hao, S. Milstein, C. Fan, M. Tipsword, K. Drew, M. Galli, K. Rhrissorrakrai, D. Drechsel, D. Koller, F. P. Roth, L. M. Iakoucheva, A. K. Dunker, R. Bonneau, K. C. Gunsalus, D. E. Hill, F. Piano, J. Tavernier, S. van den Heuvel, A. A. Hyman, and M. Vidal, “A protein domain-based interactome network for C-elegans early embryogenesis,” *Cell*, vol. 134, pp. 534–545, AUG 8 2008.

[20] D. Talavera, S. C. Lovell, and S. Whelan, “Covariation is a poor measure of molecular coevolution,” *Molecular biology and evolution*, vol. 32, no. 9, pp. 2456–2468, 2015.

[21] C. Qin and L. J. Colwell, “Power law tails in phylogenetic systems,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 4, pp. 690–695, 2018.

[22] E. Estrada, *The structure of complex networks: Theory and Applications*. Oxford University Press, 2011.

[23] J. Shendure and H. Ji, "Next-generation dna sequencing," *Nature Biotechnology*, vol. 26, no. 10, pp. 1135–1145, 2008.

[24] S. W. Lockless and R. Ranganathan, "Evolutionarily conserved pathways of energetic connectivity in protein families," *Science*, vol. 286, no. 5438, pp. 295–299, 1999.

[25] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, "Identification of direct residue contacts in protein-protein interaction by message passing," *Proc. Natl. Acad. Sci. USA*, vol. 106, no. 1, pp. 67–72, 2009.

[26] S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S.-I. Lee, and C. J. Langmead, "Learning generative models for protein fold families," *Proteins: Structure, Function, and Bioinformatics*, vol. 79, no. 4, pp. 1061–1078, 2011.

[27] D. S. Marks, T. A. Hopf, and C. Sander, "Protein structure prediction from sequence variation," *Nat. Biotech.*, vol. 30, pp. 1072–1080, NOV 2012.

[28] S. Ovchinnikov, L. Kinch, H. Park, Y. Liao, J. Pei, D. E. Kim, H. Kamisetty, N. V. Grishin, and D. Baker, "Large-scale determination of previously unsolved protein structures using evolutionary information," *eLife*, vol. 4, SEP 3 2015.

[29] S. Ovchinnikov, H. Kamisetty, and D. Baker, "Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information," *Elife*, vol. 3, 2014.

[30] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, "Protein sectors: evolutionary units of three-dimensional structure," *Cell*, vol. 138, no. 4, pp. 774–786, 2009.

[31] V. Dahirel, K. Shekhar, F. Pereyra, T. Miura, M. Artyomov, S. Talsania, T. M. Allen, M. Altfeld, M. Carrington, D. J. Irvine, B. D. Walker, and A. K. Chakraborty, "Coordinate linkage of HIV evolution reveals regions of immunological vulnerability," *Proc. Natl. Acad. Sci. USA*, vol. 108, pp. 11530–11535, JUL 12 2011.

[32] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, 2003.

[33] R. B. Cattell, "The scree test for the number of factors," *Multivariate Behavioral Research*, vol. 1, no. 2, pp. 245–276, 1966.

[34] A.-L. Barabasi, *Network Science*. Cambridge University Press, 2016.

[35] R. Weiss and A. McLean, “What have we learnt from SARS?,” *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 359, pp. 1137–1140, JUL 29 2004.

[36] R. Albert, H. Jeong, and A. Barabasi, “Error and attack tolerance of complex networks,” *Nature*, vol. 406, pp. 378–382, JUL 27 2000.

[37] K. Nguyen and D. A. Tran, *Handbook of Optimization in Complex Networks*. Springer, 2012.

[38] I. Anishchenko, S. Ovchinnikov, H. Kamisetty, and D. Baker, “Origins of coevolution between residues distant in protein 3D structures,” *Proc. Natl. Acad. Sci. USA*, vol. 114, pp. 9122–9127, AUG 22 2017.

[39] J. Baussand and A. Carbone, “A combinatorial approach to detect coevolved amino acid networks in protein families of variable divergence,” *PLoS computational biology*, vol. 5, no. 9, p. e1000488, 2009.

[40] D. Watts and S. Strogatz, “Collective dynamics of ‘small-world’ networks,” *Nature*, vol. 393, pp. 440–442, JUN 4 1998.

[41] A. Barabasi, R. Albert, and H. Jeong, “Mean-field theory for scale-free random networks,” *Physica A*, vol. 272, pp. 173–187, OCT 1 1999.

[42] R. Albert and A. Barabasi, “Statistical mechanics of complex networks,” *Rev. Mod. Phys.*, vol. 74, pp. 47–97, JAN 2002.

[43] A. F. Poon, L. W. Walker, H. Murray, R. M. McCloskey, P. R. Harrigan, and R. H. Liang, “Mapping the shapes of phylogenetic trees from human and zoonotic rna viruses,” *PLoS one*, vol. 8, no. 11, p. e78122, 2013.

[44] A. A. Hagberg, D. A. Schult, and P. J. Swart, “Exploring network structure, dynamics, and function using networkx,” in *Proceedings of the 7th Python in Science Conference* (G. Varoquaux, T. Vaught, and J. Millman, eds.), (Pasadena, CA USA), pp. 11 – 15, 2008.

[45] A. Noack, “Modularity clustering is force-directed layout,” *Physical Review E*, vol. 79, no. 2, p. 026102, 2009.

[46] A. Calderone, L. Licata, and G. Cesareni, “Virusmentha: a new resource for virus-host protein interactions,” *Nucleic acids research*, vol. 43, no. D1, pp. D588–D592, 2014.

[47] L. Dey and A. Mukhopadhyay, "Denvint: A database of protein–protein interactions between dengue virus and its hosts," *PLoS neglected tropical diseases*, vol. 11, no. 10, p. e0005879, 2017.

[48] A. Calderone, L. Castagnoli, and G. Cesareni, "Mentha: a resource for browsing integrated protein-interaction networks," *Nature methods*, vol. 10, no. 8, p. 690, 2013.

[49] M. I. Krzywinski, J. E. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra, "Circos: An information aesthetic for comparative genomics," *Genome Research*, 2009.

Chapter 3

Amino Acid Impact Factor

Abstract

Amino acid mutations in proteins are random and those mutations which are beneficial or neutral survive during the course of evolution. Conservation or co-evolution analyses are performed on the multiple sequence alignment of homologous proteins to understand how important different amino acids or groups of them are. However, these traditional analyses do not explore the directed influence of amino acid mutations, such as compensatory effects. In this work we develop a method to capture the directed evolutionary impact of one amino acid on all other amino acids, and provide a visual network representation for it. The method developed for these directed networks of inter- and intra-protein evolutionary interactions can also be used for noting the differences in amino acid evolution between the control and experimental groups. The analysis is illustrated with a few examples, where the method identifies several directed interactions of functionally critical amino acids. The impact of an amino acid is quantified as the number of amino acids that are influenced as a consequence of its mutation, and it is intended to summarize the compensatory mutations in large evolutionary sequence data sets as well as to rationally identify targets for mutagenesis when their functional significance can not be assessed using structure or conservation.

3.1 Introduction

Amino acid co-evolution analysis captures evolutionary patterns as has been presented in Chapter 2. A number of co-evolutionary methods have been developed[1–6], some methods combined the predictions from different algorithms to identify drug resistance patterns in a cohort.[7] But unlike all these methods bayesian networks show directional dependencies between amino acids, with potential implications for HIV-1 drug resistance.[8] While potentially Bayesian networks can be very powerful, the directional relations may not be robust nevertheless.[8] Furthermore, they solve inverse problem from the data which may not be as intuitive as the co-evolution networks.

Asymmetric or directional dependencies effects have been studied in other biological contexts such as in gene expression data[9] and regulatory networks.[10] However, in the context of mutations, all the efforts focused on developing symmetric co-evolution measures of amino acid pairs, and they can not suggest a directed relation between them. The idea is to capture how many amino acids j are likely to undergo a compensatory change in response to a change in amino acid i, which may be important for structure or function of the protein. These relations can be obtained between pairs of amino acids that are either within the same protein or between two interacting proteins. In order to visualize the (un)directed co-evolutionary relations, an effective tool is network representation, which has been used in metabolic interaction networks,[11] protein-protein interaction networks,[12, 13] gene regulatory networks,[14] amino acid interaction networks[15] and protein structural analysis[16, 17] as well. In this work we introduce an asymmetric measure of the directed influence of one amino acid over another amino acid from the same or another protein, use network representation for visualizing it, and illustrate the method with examples.

3.2 Methods

Sequence selection and alignment: All the sequence data other than for HIV-1 was obtained from Pfam database.[18] We used the full alignment provided by Pfam. For HIV-1 proteins, the data was obtained from Los Alamos server (https://www.hiv.lanl.gov/). Both the databases give aligned sequences. So separate sequence alignment was not performed. But the alignment was truncated to the reference protein sequence and all sequences having more than 20% gaps were eliminated from the alignment.

Master sequence: A master sequence is constructed for the MSA by using the most occurring amino acid in each position. Following the master sequence creation, each amino acid in the MSA is converted into a binary representation, denoting it by "1" if the amino acid at a given position in a sequence is the same (conserved) as the one at the same position in the master sequence, "0" otherwise. Gap is treated as 21^{st} amino acid. When gap becomes the mostly occurring amino acid, the second highest amino acid at that position is taken. While this binary classification may seem restrictive, generalizing this definition did not practically change the conclusions, as discussed later.

Directed Network: For any given pair of amino acids (i, j) two conditional probabilities are calculated:

a. $P(j = 1|i = 1) = \frac{\text{(No. of sequences with } i\text{=1 and } j\text{=1)}}{\text{(No. of sequences with } i\text{=1 and } j\text{=0 or 1)}}$, and

b. $P(j = 0|i = 0) = \frac{\text{(No. of sequences with } i\text{=0 and } j\text{=0)}}{\text{(No. of sequences with } i\text{=0 and } j\text{=0 or 1)}}$

As a probability $P(j = 1|i = 1)$ and $P(j = 0|i = 0)$ are positive numbers between 0 and 1, We consider an amino acid to be of a certain impact if both $P(j = 1|i = 1)$ and $P(j = 0|i = 0)$ are simultaneously greater than or equal to a value $0 \leq \gamma \leq 1$ which is suitably chosen depending upon the specifics of the protein and the data set used.

Statistical analysis: Statistical significance of the relation between each pair of amino acids was evaluated by a permutation test (2000 random shuffling of the columns). If p-value obtained from this statistical test was below 0.01 directional dependence was considered significant and used for further analysis.

Directed networks: If there is a directional dependence between two amino acids, treated as nodes in the network terminology, they are considered to be connected by a directed edge. The network representations for these data sets were created by displaying the directed connections.

Impact Factor: Impact factor of an amino acid i with a cut-off γ is defined as the number of amino acids j for which $P(j = 1|i = 1) \geq \gamma$ as well as $P(j = 0|i = 0) \geq \gamma$. Impact factor of amino acids is also interesting when considering inter-protein interactions. In this case as well, a similar protocol is followed. MSA of the first protein is joined with the MSA of the second protein, after matching the identities of each of the sequences and ensuring that both the proteins are from the same sample. The rest of the analysis is the same, finding the impact of residue i from the first protein, considering residue j from the second protein. The impact factor of amino acid i on the second protein is defined as the total number of all such residues j.

Dependency Factor: Similar to the impact factor, we define a dependency factor. The dependency factor of an amino acid j is the total number of amino acids which impact it with the same cut-off γ.

3.3 Results

3.3.1 Asymmetric relations and Impact calculation

In this work, we introduce directional co-evolutionary interactions among pairs of amino acids, either from the same protein or from two different proteins. We use multiple sequence alignments from homologous proteins to construct a master sequence, relative to which an amino acid in a sequence is coded "1" if it is the most occurring amino acid and "0" if it mutates to other alternatives (see **Methods** section). The dependence between amino acids i and j, schematically shown in Figure 3.1A is evaluated as follows. A definition of asymmetric dependence between amino acids at site i and j was designed using two conditional probabilities $P(j = 1|i = 1)$ and $P(j = 0|i = 0)$. The first of the two conditional probabilities reflects how amino acid i, when it is conserved, imposes conservation on j and the second reflects how a mutation in i imposes a compensatory mutation on j. When both these probabilities are greater than a predefined cut-off γ, i is considered to have an impact on j. The total number of such influences exerted by the amino acid i is defined as its impact factor. While there may be alternative creative ways to define asymmetry, the definition used here captures the directional correlations in a simple and intuitive way. The choice of γ is discussed later.

We illustrate the calculation of intra-protein impact using two proteins: Dihydrofolate reductase (DHFR) and Serine protease. DHFR plays an important role in the hydride transfer from NADPH to dihydrofolate in the reduction reaction of dihydrofolate to tetrahydrofolate. Figure 3.1B shows the two conditional probabilities discussed above for all 158 amino acids with amino acid D27 as the reference. It can be seen that on using a cut-off $\gamma = 0.8$, amino acid 27 does not have an impact on any other amino acid, while with a cut-off of 0.7, it has an impact on three other amino acids L32, D37 and F153. A partial network which shows all amino acids that are impacted by amino acid 27 is then constructed (Figure 3.1C). Another example of impact calculation was performed on serine protease, an enzyme catalyzing peptide bond cleavage. In the present work, the cut-off was used strictly, without including the few data points that may be slightly less than the cut-off. The impact factor analysis with $\gamma = 0.7$ identified 16 amino acids from DHFR and 28 amino acids from serine protease and shown on their respective three dimensional structures (Figure 3.2). The structural mapping shows that the high impact residues could be spread out everywhere, with no specific spatial preference.

3.3.2 Cut-off and Impact factor

To check the sensitivity of the analysis to cut-off parameter as well as to data curation, we repeated the analyses on serine protease. Firstly the analysis was performed by changing $\gamma = 0.7$ to 0.8. Many residues that were having impact with $\gamma = 0.7$, continued to appear with $\gamma = 0.8$ (Table B.1 in Appendix B). However, for every amino acid that appeared at $\gamma = 0.7$, $\gamma = 0.8$ reduced the number of amino acids it impacted. Thus, while the relative rank order of importance according to either of the choices of γ seems to be similar, we further explored if there is a limit to the choice of γ. In the network science terminology, the impact factor we defined is one of the centrality measures called the out-degree, which is the number of connections going outward from a given node.[19] It is obvious that as the cut-off is reduced, qualitatively number of nodes as well as the

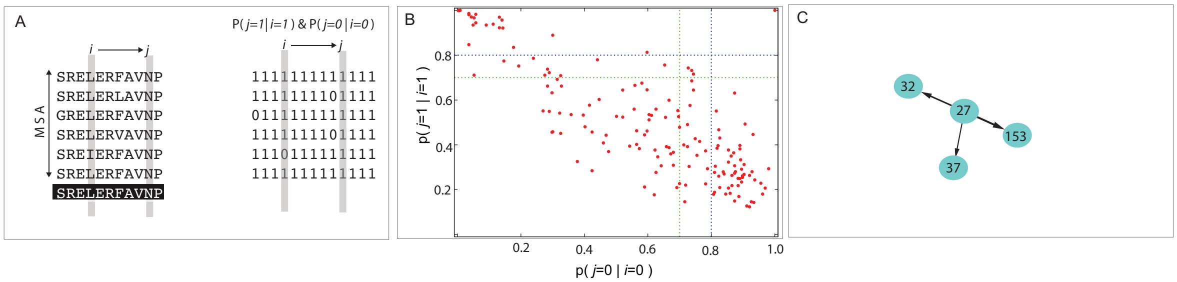

Figure 3.1: The work flow of creating directed networks. A: Schematic of the Multiple Sequence Alignment and impact calculation B: Example of the impact analysis of one of the amino acids of DHFR performed on 2303 sequences obtained from Pfam database[18] (Pfam Id: PF00186) using PDB id 3QL3 as a reference. The green and blue lines drawn at 0.7 and 0.8 represent the two cut-offs. Amino acid 27 impacts no amino acids with $\gamma = 0.8$ and 3 at $\gamma = 0.7$. The data point at $(1, 1)$ is the identity relation showing the dependence of 27 on itself. It is not used in the analyses. C: Partial network that was constructed for the impact of amino acid 27 and $\gamma = 0.7$.

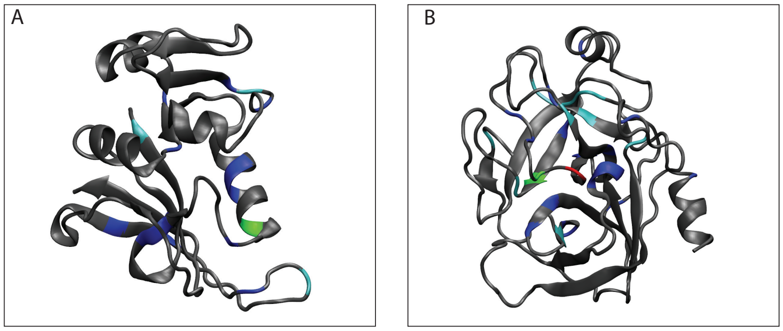

Figure 3.2: Amino acid residues with non-zero impact factor represented on the three dimensional structures of proteins: **A**. DHFR. **B**. serine protease. **Impact factor** (amino acids) for DHFR is: **3** (27), **2** (3, 57, 146), **1** (13, 14, 22, 31, 32, 55, 58, 90, 95, 135, 138, 149) and for serine protease is **8** (196), **3** (140, 194), **2** (19, 34, 102, 142, 182, 183, 184, 216, 228), **1** (29, 32, 40, 42, 57, 58, 100, 122, 136, 168, 189, 191, 201, 211, 226, 237). The coloring convention for PDBs is: Impact 0 - gray, Impact 1 - blue, Impact 2 - cyan, Impact 3 - green, Impact 8 - red.

number of outward going connections increase. We make a statistical comparison at the complete network level by using node-degree distribution,[19] which plots number n vs. the number of nodes with n outward going connections. The node-degree distributions were analysed with different choices of γ for serine protease and DHFR (Figures B.1 and B.2 in Appendix B). The node degree distribution with power-law and poissonian distributions are used to differentiate between ordered and random nature of network connections.[19] InFigures B.1 and B.2 of Appendix B, one can see that below a certain cut-off the graphs transition from power-law behavior towards poissonian distribution, suggesting a transition to random-networks. The choice of cut-off can thus be limited by these node-degree distributions to avoid the system-level random connections.

When the number of sequences were halved, the master sequence itself can change in principle, especially if a site has a conservation less than 30% or where there are two residues with comparable frequency of occurrence. Among all the proteins we studied, even though there were a few changes in the master sequences when the data set was randomly halved, there were no changes in the amino acid interaction networks, except in the case of Phosphoglycerate kinase (PGK). For PGK one residue position which had appeared in the network had many connections and were not retained when the number of sequences were changed.

3.3.3 Directed networks and Functional relevance

Distal mutation in DHFR: Using the present approach we summarize the compensatory mutations seen in the 2303 DHFR homologous sequences from the Pfam database (Pfam Id :PF00186). Performing the impact factor analysis in DHFR shows that 16 amino acids have impact with $\gamma = 0.7$. 21 connections were identified using the conditional probability criteria described in the **Methods** section and all of them except one were found to have p-value less than 0.01. The residues obtained with $\gamma = 0.7$ are shown on the three dimensional structure of DHFR labeled with the color coding corresponding to impact factor (Figure 3.2A). All the identified directed interactions are shown in Figure 3.3A as a network representation. The residues near to the folate binding pocket are found to have impact on each other. Also the catalytic residue F31 has an impact on catalytic residue I94. More interestingly the mutation at the residue position V13 has an impact on residue G121 which may be essential for maintaining the correlated dynamics between Met20 loop and the region near G121 and hence the catalytic activity.[20, 21] Also it is notable that most of the interacting pair of residues identified in this way are near in structure even though are far in sequence. The residues belonging to each of the disconnected components of the network have comparable conservation.

Catalytic residues in Serine Protease: Impact analysis on 14659 sequences obtained from Pfam (Pfam Id:PF00089) homologous to the 223 residue long serine protease shows that there are 28 residues with non-zero impact factor at $\gamma = 0.7$, 11 with $\gamma = 0.8$ and 3 with $\gamma = 0.9$. Amino acids G216, G226, D189 and V183 which were functionally associated with the rates of catalysis experimentally and in the sector analysis (red sector) are captured with this impact analysis.[22] In the case of serine protease also most of the residue pairs identified are near in structure as clear from the network representation of the interactions (Figure 3.3B). Most interestingly the catalytic triad (H57, D102 and S195) are found to have impact on each other. Also the co-evolving disulphide bond pair C42 and C58 plays important role in catalysis by optimally positioning H57 of the catalytic triad.[23]

Compensatory mutations in HIV protease and Gag: HIV protease cleaves the Gag and Gag-Pol polyproteins into individual proteins and hence is vital for the viral maturation. Many of the drugs for HIV target protease. The impact factor analysis on 2550 HIV-1 subtype B protease sequences downloaded from the Los Alamos HIV database(http://www.hiv.lanl.gov/) identified 28 compensatory mutation pairs with $\gamma = 0.8$. Residue L76 which is located near to the active site cavity is found to have a high impact factor of 6. The compensatory mutation pair V32-M46 has previously been observed experimentally[24] which showed that the reduced replication capacity of the virus due to the mutation V32I is restored by mutation M46I.

HIV virus gains resistance against the protease inhibitors on accumulation of multiple mutations not only in protease but also in the Gag polyprotein[25–27]. Our Gag-Pol inter-protein impact factor analysis with $\gamma = 0.9$ captured some of the possible compensatory mutations: positions near to the cleavage sites in Gag - A431, G381, P133 acting as compensatory mutations for the protease mutations at L76, L38 and G52 respectively. Changing the cut-off from $\gamma = 0.9$ to 0.8 resulted in the intra-protein connections increasing from 901 to 1336 and comparably, the inter-protein connections increasing from 266 to 521, highlighting the number of inter-protein compensatory effects.

Compensatory Mutation in PGK: Phosphoglycerate kinase (PGK) is involved in the ATP generating step of glycolytic pathway: the reversible reaction of 1,3-bisphosphoglycerate and ADP to 3-phosphoglycerate and ATP. The catalytic residues of PGK is highly conserved across different species. But it is observed that the residue 219 of PGK which is crucial in the dynamics facilitating catalysis is lysine in all Eukaryotes and Bacteria where as it is threonine or serine in *Archaea*.[28] The loss of catalytic activity due to this mutation (K219S) is found to have been restored by compensatory mutations at the positions 239 and 403.[28] Through our impact factor analysis of the Pfam family PF00162 with $\gamma = 0.8$ we could capture the compensatory mutation at the site 403.

3.4 Discussion

3.4.1 Directed co-evolution

The present work develops two principles: directed co-evolutionary relationships between amino acids and a quantification of it by counting the number of such dependencies. Amino acids in the primary chain of the protein contribute to its structural stability or function and mutations of these amino acids are differentially tolerated. At a simplistic level, considering the tolerance to the variations in amino acids and/or its neighbors, they

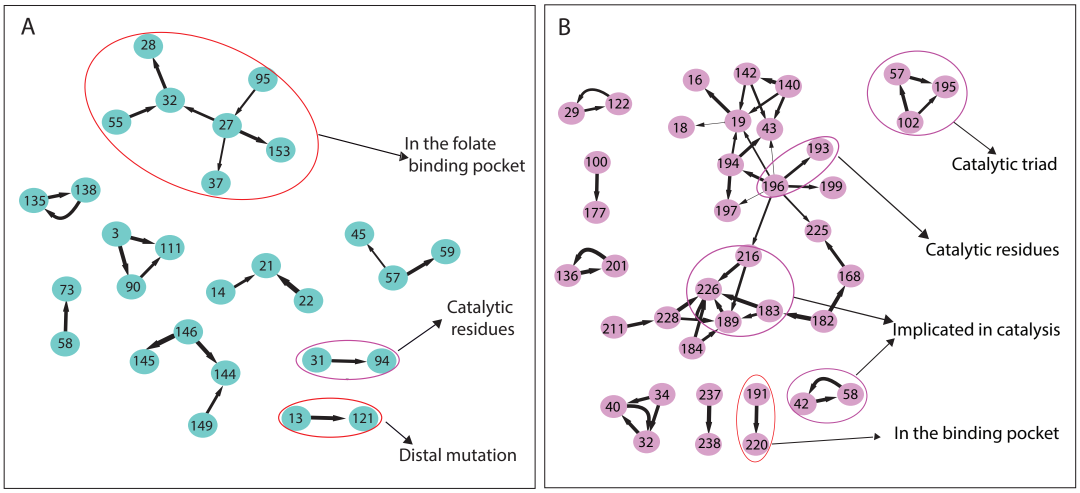

Figure 3.3: Directed networks and their functional relevance. Residue networks for **A**. DHFR (PDB Id:3QL3) and **B**. Serine protease (PDB Id:3TGI). The direction of the arrow shows is in the direction of impact. The thickness of the arrows is proportional to $1/r$ where r is the distance between pair of amino acids in the crystal structure. The functional annotation of the amino acids inferred from literature is shown as well.

may be grouped as: (i) absolutely essential and hence can not mutated, (ii) essential but may tolerate certain substitutions, (iii) essential and tolerate substitutions with suitable compensatory mutations and (iv) not-essential. The present method for identifying directed co-evolutionary relation is mainly to address the amino acids in group (iii). Groups (i) and (ii) are mostly captured by conservation analysis. In fact, if an amino acid is so essential that it was never replaced or evolved among the sequences studied, it will not appear in any co-evolutionary analysis. Further, the directed co-evolution relation developed should not be construed as a description of causal relationships. It represents a statistical summary of the interdependencies among different amino acids while studying large sets of sequence data to identify possible compensatory effects.

In general, when the conditional probabilities of the mutation of one amino acid relative to all other amino acids are studied, such as in Figure 3.1, the number of relations are few. When the positions i and j are uncorrelated, $P(j = 1|i = 1) = P(j = 1)$. Similarly, $P(j = 0|i = 0) = P(j = 0)$, which is identically $1 - P(j = 1)$. So, all the amino acid mutations that are uncorrelated scatter in the anti-diagonal way, as seen in Figure 3.1B. With a relatively high cut-off γ, only a few amino acid relations appear in the zone of interest, which is on the top-right corner. This could be seen from the average number of impact relations that were identified after seeking a significance level $p < 0.01$ (Table B.2 in Appendix B). Although we look for directed relations between amino acids, at times the relations may be reciprocated. These reciprocal relations which are signature of co-evolution are incidental, but the focus of the present analysis remains to be the relation between a specific pair, one specific direction at a time.

3.4.2 Relation to conservation and dependency

Functional residues tend to have a higher conservation. Recent studies suggest that most of the information that is contained in the important amino acids identified using SCA is reflected by their conservation.[29] However, under certain conditions, a mutation at these positions can be compensated by changes in other amino acids. We studied the relation of the amino acid impact factors obtained in our calculations to their respective conservation scores. For DHFR and serine protease (Figure 3.4A) as well as for HIV-1 protease and reverse transcriptase (Figure B.3 in Appendix B), we see that the impact factor can not be directly inferred from conservation data alone, and as such it is not a trivial repetition of conservation. Amino acids with low conservation can have a high impact and vice-versa. The spread in conservation for the high impact residues is much broader for DHFR and serine protease as they are obtained from across the species (Figure 3.4A), compared to that in HIV-1 protease and reverse transcriptase which are obtained from the polymorphisms in the cohort (Figure B.3 in Appendix B). Despite these characteristic differences expected in the conservation patterns in these viral and non-viral proteins, the conclusion about lack of its correlation with impact could be seen in both cases.

Figure 3.4B shows impact versus dependency for DHFR and serine protease. There are several amino acids which have both high impact and dependency. This counter-intuitive behavior comes from some reciprocal relations. It is also possible that these amino acids are intermediates in the interaction network. However for many of these amino acids the higher the impact, the lesser the dependency, which highlights the importance of looking at directed compensatory effects as well rather than co-evolutionary measures alone.

3.4.3 Three-state model

One apparent limitation that arises from the above analyses is that they use a binary-model: at any position the amino acid corresponds to the one in master sequence or not. Practically, in the data sets we used, we saw a few discrete scenarios where two dominant polymorphisms occurred with comparable frequencies. Hence, without complicating for the theoretical possibility of large number of polymorphisms, we performed a three-state model as the next step towards generalizing our model. In this model we considered residues which occur with a frequency more than 35% at a position to be distinct states. Since there can be at most two states which have a frequency more than 35%, the amino acid code in a sequence is replaced by "1" or "2" depending on the polymorphic state

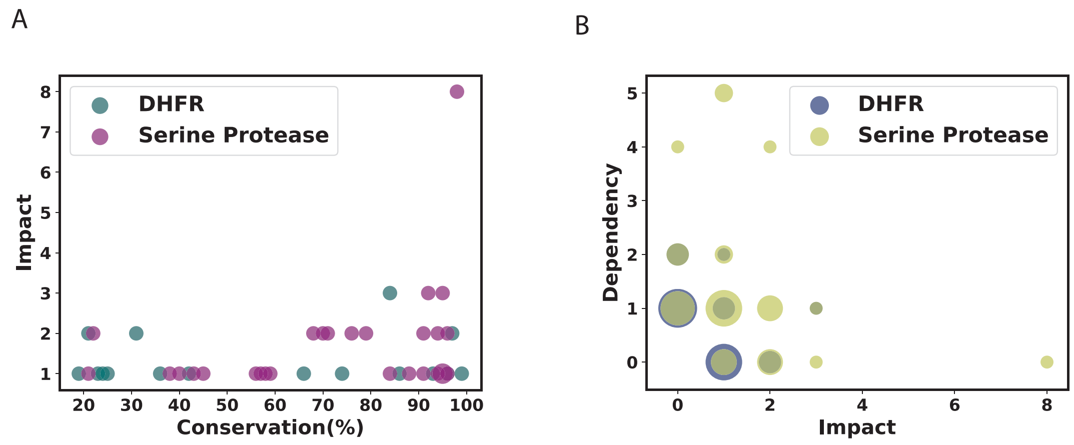

Figure 3.4: Comparisons of impact with other measures: **A**. Impact vs. conservation shows that the impact does not trivially repeat the same information contained in conservation. **B**. Impact vs. dependency shows again in addition to the expected negative correlation between the two, there are several deviations from it. Impact was calculated with $\gamma = 0.7$. (Size of the marker shows the density of points at that position)

and "0" if it did not belong to either. So the conditional probabilities to be satisfied for position i to have an impact on position j is:
$P(j = 1|i = 1) \geq \gamma, P(j = 2|i = 2) \geq \gamma$ and $P(j = 0|i = 0) \geq \gamma$
or $P(j = 2|i = 1) \geq \gamma, P(j = 1|i = 2) \geq \gamma$ and $P(j = 0|i = 0) \geq \gamma$

When we repeated the analysis with $\gamma = 0.7$, the new three-state definition was relevant only to a few amino acids: 1, 6 and 14 amino acids from serine protease, DHFR and PGK respectively. However despite this three state generalisation, these positions from serine protease and DHFR did not appear in the directed co-evolution network. But in the case of PGK, 2 out of 14 had appeared in the network when the binary model is used and the connections involving these residues do not appear when this new definition is used. Thus in the spirit of the inclusive definition of identifying important residues, a more restrictive binary-state definition with slightly more network connections seems suited for the analysis.

3.4.4 Significance of impact analysis

Using the directed network analysis, several critical amino acids with functional significance could be identified as discussed in the previous section. The perfectly conserved amino acids that never evolve, which are very likely with high functional significance, do not appear in the analysis by the nature of the definition. Those amino acids could anyway be identified using the conservation analysis. The significance of the present analysis should thus be seen as one that identifies amino acids which are likely to have functional repercussions unless compensated, and are not obvious from the standard conservation analyses. Thus the present analysis is to be treated as an inclusive analysis, rather than a comprehensive one, to suggest which amino acids should be included into further analyses - experimental or theoretical. In that sense, the residues requiring the most number of compensatory mutations, may be considered as the significant ones in the analysis. This knowledge may be useful while studying intra- or inter-protein amino acid correlations among large sets of evolutionary or cohort data, and for forming a rational basis for performing site directed mutagenesis experiments. The meaning of the numeric value of the impact factor itself may not be obvious, especially when comparing analyses across two different protein families. However, within one protein family the impact factor rank-orders the different amino acids by summarising the evolutionary data and prioritising them for mutagenesis experiments.

3.4.5 Resistance models

The notion and definition of directed networks can be generalised to other cases. For example, in analysing the clinical data of the bacterial strains from the group of patients who respond (sensitive) to a drug versus those that do not (resistant), the same principles may be used. Resistance to antibiotics poses a severe public health problem, and usually there is a strong correlation between drug usage patterns in a cohort or a geographic region[30] and the development of bacterial resistance. Mutations of amino acids from critical bacterial or viral proteins that are the targets in drug design, may lead to a fitness advantage. However, these mutations may have to be compensated by other mutations in other sites in the same or other proteins.[31, 32]. For example, in ribosomal protein S12, which is a usual drug target, K42N mutation may be compensated by as many as 35 mutations from both the same protein as well as from others[31]. It is important to identify these compensations that go on in the drug-resistant cohort from the perspective of avoiding problems with secondary drug resistance.

With such background about cohorts and compensatory mutations, one might design questions such as - which are the amino acids j that had a compensatory mutation ($j = 0$) in the resistant group when a drug targeting amino acid i is used. These mutations in j contribute to a structural or functional compensation for a mutation in i that made it drug resistant, rather than requiring a reversion of the mutation in i.[31] Thus comparing the resistant and sensitive cohorts one can evaluate if the conditional probabilities $P_{resistant}(j = 0|i = 0)$ and $P_{sensitive}(j = 1|i = 1)$ exceed a threshold, γ. The method is equally applicable when i and j are from the same protein or from two proteins whose sequences are juxtaposed to perform similar analysis. This analysis is a simpler alternative to the Bayesian analyses that are sometimes used for the specific mutations in the drug-resistance group.[8]

Directed co-evolutionary relationships can be useful either from the protein design or drug design perspective. Considering the compensatory effects, one may plan to add simultaneous mutations along with mutations that contribute to the specific functional gain or design combination therapies such that the primary group of amino acids targeted by the drug, as well as those that undergo consequent mutations are simultaneously targeted.

3.5 Conclusions

We introduced a way to measure and visualise the directed influence of amino acids on one another. The directed influence network summarizes the compensatory mutations under functional constraints in response to changes of key amino acids in homologous sequences. We demonstrate the utility of the method using evolutionary sequences from a few proteins. The principal results seem to be unaffected by changes in parameters and identify effects from compensation to distal mutations, as well as the binding pocket and catalytic residues. The simple and intuitive definition of the directional impact of amino acid interactions can bring a new perspective to the field that had so far relied on symmetric co-evolution.

Bibliography

[1] S. W. Lockless and R. Ranganathan, "Evolutionarily conserved pathways of energetic connectivity in protein families," *Science*, vol. 286, no. 5438, pp. 295–299, 1999.

[2] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, "Identification of direct residue contacts in protein–protein interaction by message passing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 1, pp. 67–72, 2009.

[3] J. Baussand and A. Carbone, "A Combinatorial Approach to Detect Coevolved Amino Acid Networks in Protein Families of Variable Divergence," *PLoS Computational Biology*, vol. 5, SEP 2009.

[4] S. H. Ackerman, E. R. Tillier, and D. L. Gatti, "Accurate simulation and detection of coevolution signals in multiple sequence alignments," *PLoS One*, vol. 7, p. e47108, 2012.

[5] S. Bremm, T. Schreck, P. Boba, S. Held, and K. Hamacher, "Computing and visually analyzing mutual information in molecular co-evolution," *BMC bioinformatics*, vol. 11, no. 1, p. 330, 2010.

[6] D. de Juan, F. Pazos, and A. Valencia, "Emerging methods in protein co-evolution," *Nature Reviews Genetics*, vol. 14, pp. 249–261, APR 2013.

[7] G. Li, K. Theys, J. Verheyen, A.-C. Pineda-Peña, R. Khouri, S. Piampongsant, M. Eusébio, J. Ramon, and A.-M. Vandamme, "A new ensemble coevolution system for detecting hiv-1 protein coevolution," *Biology Direct*, vol. 10, no. 1, p. 1, 2015.

[8] K. Deforche, T. Silander, R. Camacho, Z. . Grossman, M. A. Soares, K. Van Laethem, R. Kantor, Y. Moreau, A. M. Vandamme, and non B Workgroup, "Analysis of HIV-1 pol sequences using Bayesian Networks: implications for drug resistance," *Bioinformatics*, vol. 22, pp. 2975–2979, DEC 15 2006.

[9] N. Guelzim, S. Bottani, P. Bourgine, and F. Képès, "Topological and causal structure of the yeast transcriptional regulatory network," *Nature Genetics*, vol. 31, no. 1, p. 60, 2002.

[10] R. Opgen-Rhein and K. Strimmer, "From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data," *BMC Systems Biology*, vol. 1, no. 1, p. 37, 2007.

[11] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of escherichia coli," *Nature genetics*, vol. 31, no. 1, pp. 64–68, 2002.

[12] J. Wang, S. Rao, J. Chu, X. Shen, D. N. Levasseur, T. W. Theunissen, and S. H. Orkin, "A protein interaction network for pluripotency of embryonic stem cells," *Nature*, vol. 444, no. 7117, pp. 364–368, 2006.

[13] G. Kar, A. Gursoy, and O. Keskin, "Human cancer protein-protein interaction network: a structural perspective," *PLoS Computational Biology*, vol. 5, no. 12, p. e1000601, 2009.

[14] G. Karlebach and R. Shamir, "Modelling and analysis of gene regulatory networks," *Nature Reviews Molecular Cell Biology*, vol. 9, no. 10, pp. 770–780, 2008.

[15] R. Aurora, M. J. Donlin, N. A. Cannon, J. E. Tavis, V.-C. S. Group, *et al.*, "Genome-wide hepatitis c virus amino acid covariance networks can predict response to antiviral therapy in humans," *The Journal of Clinical Investigation*, vol. 119, no. 1, p. 225, 2009.

[16] G. Amitai, A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger, and S. Pietrokovski, "Network analysis of protein structures identifies functional residues," *Journal of Molecular Biology*, vol. 344, no. 4, pp. 1135–1146, 2004.

[17] K. Brinda and S. Vishveshwara, "A network representation of protein structures: implications for protein stability," *Biophysical Journal*, vol. 89, no. 6, pp. 4159–4170, 2005.

[18] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, *et al.*, "The pfam protein families database: towards a more sustainable future," *Nucleic Acids Research*, vol. 44, no. D1, pp. D279–D285, 2016.

[19] A.-L. Barabasi, *Network Science.* Cambridge University Press, 2016.

[20] T. H. Rod, J. L. Radkiewicz, and C. L. Brooks, "Correlated motion and the effect of distal mutations in dihydrofolate reductase," *Proceedings of the National Academy of Sciences*, vol. 100, no. 12, pp. 6980–6985, 2003.

[21] C. E. Cameron and S. J. Benkovic, "Evidence for a functional role of the dynamics of glycine-121 of escherichia coli dihydrofolate reductase obtained from kinetic analysis of a site-directed mutant," *Biochemistry*, vol. 36, no. 50, pp. 15792–15800, 1997. PMID: 9398309.

[22] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, "Protein sectors: evolutionary units of three-dimensional structure," *Cell*, vol. 138, no. 4, pp. 774–786, 2009.

[23] T. T. Baird, W. D. Wright, and C. S. Craik, "Conversion of trypsin to a functional threonine protease," *Protein Science*, vol. 15, no. 6, pp. 1229–1238, 2006.

[24] A. M. Borman, S. Paulous, and F. Clavel, "Resistance of human immunodeficiency virus type 1 to protease inhibitors: selection of resistance mutations in the presence and absence of the drug," *Journal of General Virology*, vol. 77, no. 3, pp. 419–426, 1996.

[25] M. F. Maguire, R. Guinea, P. Griffin, S. Macmanus, R. C. Elston, J. Wolfram, N. Richards, M. H. Hanlon, D. J. Porter, T. Wrin, *et al.*, "Changes in human immunodeficiency virus type 1 gag at positions l449 and p453 are linked to i50v protease mutants in vivo and cause reduction of sensitivity to amprenavir and improved viral fitness in vitro," *Journal of Virology*, vol. 76, no. 15, pp. 7398–7406, 2002.

[26] H. Gatanaga, Y. Suzuki, H. Tsang, K. Yoshimura, M. F. Kavlick, K. Nagashima, R. J. Gorelick, S. Mardy, C. Tang, M. F. Summers, *et al.*, "Amino acid substitutions in gag protein at non-cleavage sites are indispensable for the development of a high multitude of hiv-1 resistance against protease inhibitors," *Journal of Biological Chemistry*, vol. 277, no. 8, pp. 5952–5961, 2002.

[27] L. Doyon, G. Croteau, D. Thibeault, F. Poulin, L. Pilote, and D. Lamarre, "Second locus involved in human immunodeficiency virus type 1 resistance to protease inhibitors.," *Journal of Virology*, vol. 70, no. 6, pp. 3763–3769, 1996.

[28] A. Wellner, M. R. Gurevich, and D. S. Tawfik, "Mechanisms of protein sequence divergence and incompatibility," *PLoS Genetics*, vol. 9, no. 7, p. e1003665, 2013.

[29] T. Teşileanu, L. J. Colwell, and S. Leibler, "Protein sectors: statistical coupling analysis versus conservation," *PLoS Computational Biology*, vol. 11, no. 2, p. e1004091, 2015.

[30] H. Goossens, M. Ferech, R. Stichele, M. Elseviers, and E. P. Grp, "Outpatient antibiotic use in Europe and association with resistance: a cross-national database study.," *The Lancet*, vol. 365, pp. 579–587, FEB 12 2005.

[31] S. Maisnier-Patin and D. Andersson, "Adaptation to the deleterious effects of antimicrobial drug resistance mutations by compensatory evolution," *Research in Microbiology*, vol. 155, pp. 360–369, JUN 2004.

[32] B. Levin, V. Perrot, and N. Walker, "Compensatory mutations, antibiotic resistance and the population genetics of adaptive evolution in bacteria," *Genetics*, vol. 154, pp. 985–997, MAR 2000.

Chapter 4

Correlations from Structure, Sequence and Dynamics are Complementary Rather than Synonymous

Abstract

Amino acid interaction brings structural stability as well as structural changes according to the environment required for the function of the protein. There have been different approaches based on sequence, structure or dynamics of the protein for identifying such important interactions. In this work we compare the interactions and amino acids identified by each of these methods for the proteins rat trypsin protease and dihydrofolate reductase. We found that the overlap of connections that are ranked top based on the strength in all three analyses are few while structure and dynamics share 40% of pairs in common even when as little as 100 connections are chosen. Though it is well established that the sequence information can be used to predict protein structure and dynamics of the protein is highly correlated with the structure, we find that there is unique information in each of these and the complete understanding of all important interactions requires analyses based on sequence, structure and dynamics of the protein.

4.1 Introduction

Proteins perform several critical cellular functions. Experimental,[1, 2] theoretical[3, 4] and computational[5, 6] efforts over decades have elucidated several aspects of how proteins fold and function. Despite this rich landscape of studies, it is still hard to predict how proteins function. Even in the background of new detailed studies such as large scale mutational scans,[7] predicting how and why a given mutation affects the function and the pathways of effect propagation has not been easy. The emphasis is thus on functionally critical amino acids, proximal or distal amino acids that influence these functional amino acids and the pathways that connect them.

There are evidences to show that the dynamic nature of proteins is required for their function.[8] However the dynamics centers around the native structure of the protein which is uniquely determined by its sequence. Thus in principle any of the three descriptors- sequence, structure and dynamics should be sufficient to understand the protein function or the mutational effects, which are important from basic biology and protein design perspectives.

The simplest sequence based analysis uses conservation of an amino acid seen in multiple sequence alignments of homologous proteins.[9] Pair-wise co-evolution relations which measure the statistical chance that an amino acid mutates whenever there is a change in the others became a way for identifying effects which cannot be easily interpreted from structure.[10–12] These analyses look for amino acid positions that evolve together hence effects of change at one position compensating for the effects of amino acid substitution at another position. The co-evolving group of amino acids termed as sectors are found to be associated with specific functions related to catalysis and stability.[13]

Some of the structure based analyses explore how an amino acid may be critical for the protein function because it is a functionally important one such functional amino acid, or interacts through a pathway of residues connecting the two.[14] Commute time which is the time required for a signal to travel to and fro between two amino acids is used for quantifying this speed of communication and to identify the pathways between them. Residues though far from the catalytic sites can affect catalysis if has indirect interaction with the catalytic site. Commute time analysis can identify even these allosteric sites along with the structural contacts.

Dynamics is another crucial factor that determines the function of the protein. It has been observed that correlated motion of groups of residues bring out structural changes which are functionally important. For example binding and unbinding of ligands require simultaneous movement of a group of residues.[15] Understanding this correlated motion of residues will help in figuring out the allosteric pathway and hence in designing better drugs. Molecular dynamics simulation studies using elastic network or all-atom models have been used for this purpose. Dynamic cross correlation,[16] inter-residue distance fluctuation and other information theoretic approaches[17, 18] are proposed to capture residues with correlated dynamics. The concept of "dynamics sectors" similar to as that in the case of sequences where amino acids are grouped based on the dynamics also exist.[19, 20]

Each of the above mentioned approaches based on sequence,[13] structure[14] and dynamics have been used individually to identify important amino acids or interactions. The choice of the method was mostly dictated by the training of the individual scientists. Irrespective of the choice of the approach, more often than not, a strong signal in any of these approaches was used to justify the observed allosteric effects. Since sequence, structure and dynamics are related, it is implicitly believed that these three approaches predict the same critical amino acids, without any systematic validation. To the best of our knowledge, these approaches have not been examined for false positives, with a goal of building predictability. All studies lacked a comprehensive and detailed comparison of methods based on sequence, structure and dynamics. In this study we attempt to compare the results from analyses based on each of the sequence, structure and dynamics aspects for two proteins, serine protease and dihydrofolate reductase (DHFR), and highlight both the similarities and differences. Serine protease and DHFR were chosen because of the availability of mutational as well as computational analysis for comparisons.

4.2 Methods

Sequence analysis: The multiple sequence alignment for rat serine protease and E. coli. dihydrofolate reductase were obtained from Pfam database[21] with pfam IDs PF00089 and PF00186 respectively. As the full alignment had more than 43,000 sequences the analysis was performed using randomly selected 10,000 sequences which had an identity more than 20%. All sequences with a gap frequency more than 20% compared to the reference sequence were removed from the alignment. The positions with gap frequency more than 20% were not included in the analysis. The alignment was then binarized by assigning 1 to the mostly conserved amino acid at a position and 0 to any other amino acid. The Statistical Coupling (SC) matrix was calculated following Halabi et al.[13] using this alignment. The top most eigen component as well as the components with eigenvalues smaller than that of the SC matrix calculated for a random alignment were removed. This cleaned SC matrix scores were used for all further analyses.

Commute time analysis: In this analysis the signal propagation in proteins is modeled as a discrete-time, discrete-state Markov process in which information is transferred across the network of amino acids as obtained from the 3D structure of the protein. The strength of communication between two amino acids A and B is then quantified based on the number of steps taken for signal to travel from A to B and then back to A which is called as commute time.[14] In order to calculate this, the protein structure is considered as a network of n nodes, where n is the number of amino acids in the protein. The strength of interaction, also called affinity between two residues is defined as $a_{ij} = N_{ij}/\sqrt{N_i N_j}$ where N_{ij} is the number of atom-atom contacts within 4 Å between residues i and j and N_i and N_j are the number of heavy atoms in residues i and j. The probability of transferring information present at residue j to residue i is calculated as $m_{ij} = d_j^{-1} a_{ij}$ where $d_j = \sum_{i=1}^{n} a_{ij}$. $M = m_{ij}$ is called the Markov transition matrix. The hitting time or the average number of steps taken for signal to travel from residue j to i is then determined by solving the self consistent set of equations, $H_{ij} = 1 + \sum_{k=1}^{n} H_{jk} m_{ki}$. Commute time is then calculated as $C_{ij} = H_{ij} + H_{ji}$. For more details see Chennubhotla et al.[14]

Molecular dynamics simulations: Molecular dynamics simulations were performed for both proteins using the PDB IDs 3TGI (serine protease) and 3QL3 (DHFR) as the starting structures and GROMACS 5.1.4[22]. The structures were energy-minimized using steepest descent algorithm followed by NVT and NPT simulations each of 1ns long with 2 fs as the time step. All bonds involving hydrogen were constrained using LINCS algorithm. Velocity rescaling was used as thremostats in NVT and NPT simulation. For equilibration Parrinello-Rahman barostat was used. The NPT equilibration was followed by the 100 ns production run in the NPT ensemble at 300 K and 1 bar pressure. The trajectory was saved at every 1 ps. This 100 ns trajectory was used for the dynamics related calculations. We understand that this trajectory is very short as compared to the timescales of the functional dynamics of proteins which is of the order of milliseconds to seconds. But to fully sample this dynamics, the trajectories have to be a few orders of magnitude longer than that, which apart from our limitations on computational time, may also raise questions on the suitability of force fields. While acknowledging these limitations, in this work, we perform simulations of 100 ns, as is the common practice while searching for dynamical cross-correlations, and analyse this data of equilibrium fluctuations around the native structure.

Inter-residue distance fluctuations: The inter-residue distance fluctuation was calculated as $\langle |\Delta r_{ij}|^2 \rangle = \langle |\Delta r_j - \Delta r_i|^2 \rangle$ where $\Delta r_i = r_i - \langle r_i \rangle_t$ and r_i and r_j are the position vectors of the C_α carbon atoms of residues i and j respectively. $\langle r_i \rangle_t$ is the average position of residue i over the 100 ns trajectory.

Node strength from networks: The node strength which is the sum of all edge weights that a node is involved in were calculated using Networkx module of Python. For structure and dynamics related networks, inverse of commute time and inverse of inter-residue distance fluctuation respectively were used as the weight of interaction. Whereas for sequence, the co-evolution values were used as weight.

Network representation: Cytoscape was used to create all network representations.

4.3 Results

4.3.1 Pair-wise interactions

We performed pair-wise amino acid co-evolution based on Statistical Coupling Analysis (SCA),[13] commute time analysis[14] using the structure, and inter-residue distance fluctuations from molecular dynamics simulation respectively. Qualitatively two amino acids are believed to be interacting when the correlations in SCA are higher or the commute time or distance fluctuation matrices is lower. The SC, commute time and inter-residue distance fluctuation matrices calculated (**Methods**) for serine protease and DHFR and are given in Figure 4.1A,4.1C,4.1E and 4.1B,4.1D,4.1F respectively. The correlation between these matrices are given in Table C.1 of Appendix C. From this matrix representation, it can be seen that the commute-time and fluctuation matrices have similar patterns as these are expected to be linearly related in the elastic network framework of proteins. A matrix representation such as the one in Figure 4.1 is helpful for visualizing the similarity pattern as well as to quickly note if a specific interaction $i-j$ is strong. For other analyses we translated the interaction matrices into networks by using the matrix elements as adjacency factors.

4.3.2 Hubs of interaction

In the network representation we performed analyses beyond noting the individual pairwise relations. It is possible that some residues interact with many others, acting as the hubs of interaction, thus assuming a central and critical role in the protein. In order to identify these residues, the residues were sorted according to the total node interaction strength in the complete residue-residue interaction network, and the ones having high interaction strength were selected. This procedure was repeated for all three interaction networks. In Figure 4.2 the node strength of all residues in each of the networks is represented as different node attributes in the contact network of the protein. The residues in the core of the protein has higher node strengths more than the ones on the surface. This particular representation which serves as a two-dimensional projection of the protein structure helps in visualizing the layout of different important amino acids.

4.3.3 Comparison of highly interacting residues

We investigated how similar the important nodes selected from the different approaches are. Figure 4.3 shows the number of residues that are common between important residues of structure, sequence and dynamics networks when the top 20 nodes from each network are selected. As shown in Figure4.4 the convergence of the three approaches is poor when less than 50 amino acids are chosen from each of the methods. It can also be seen

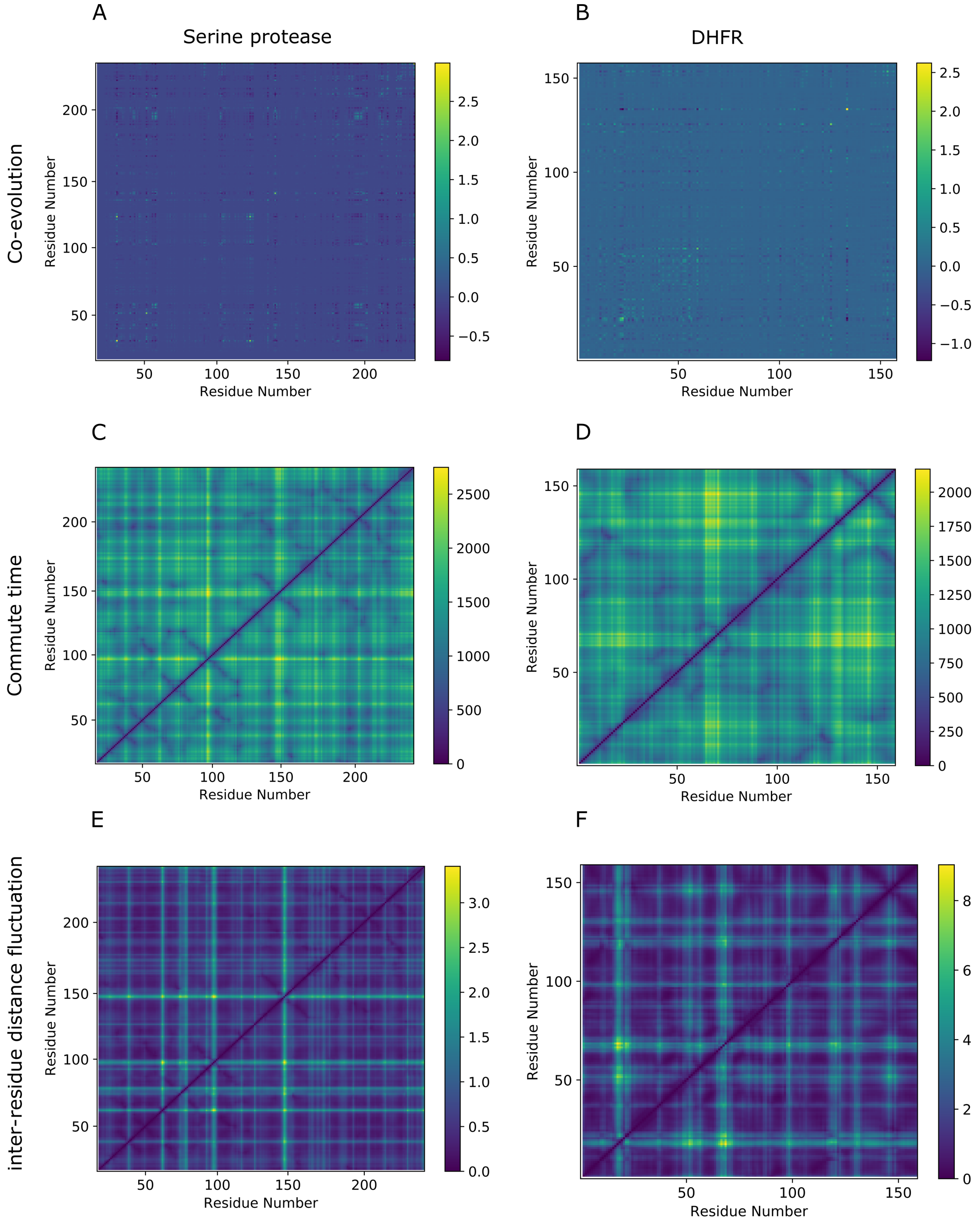

Figure 4.1: Pair-wise amino acid interactions obtained by analyses based on sequence, structure and dynamics - A,C,E are of serine protease and B,D,F are of DHFR. Amino acid co-evolution (A and B) from the multiple sequence alignment (MSA) of the protein was calculated using SCA protocol. Strength of interaction based on structure was quantified as the commute-time (D and E) between amino acid pairs and fluctuations in the inter-residue distances (F and G) calculated from the all-atom MD simulation trajectory was used to determine amino acids with correlated dynamics. Lower commute-time and lower inter-residue distance fluctuation correspond to stronger interaction.

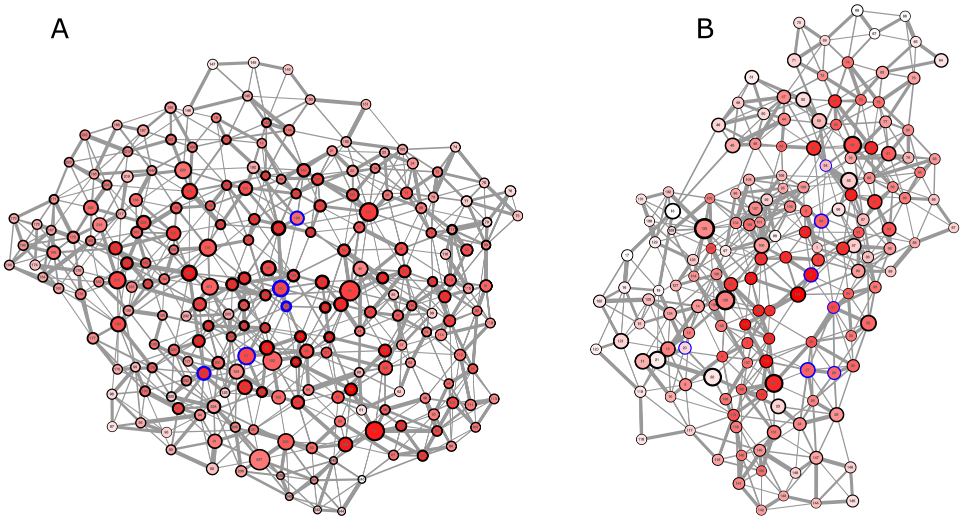

Figure 4.2: Depiction of centrality measures for all amino acids on the contact network of the proteins (A) Serine protease (B) DHFR. The structural contacts were determined from the PDB structure of the protein. All residue pairs having contacts (atoms within 4Å) are shown as connected with the edge thickness proportional to the number of contacts. The size, color and border thickness of the nodes represent the node weights in the co-evolution, commute time and inter-residue distance fluctuation matrices respectively. The catalytic residues are highlighted with blue borders.

that the structure and dynamics has higher overlap, but this convergence is also far from being ideal. Varying the number of nodes chosen, the common nodes increases and this variation is higher compared to the case of random selection.

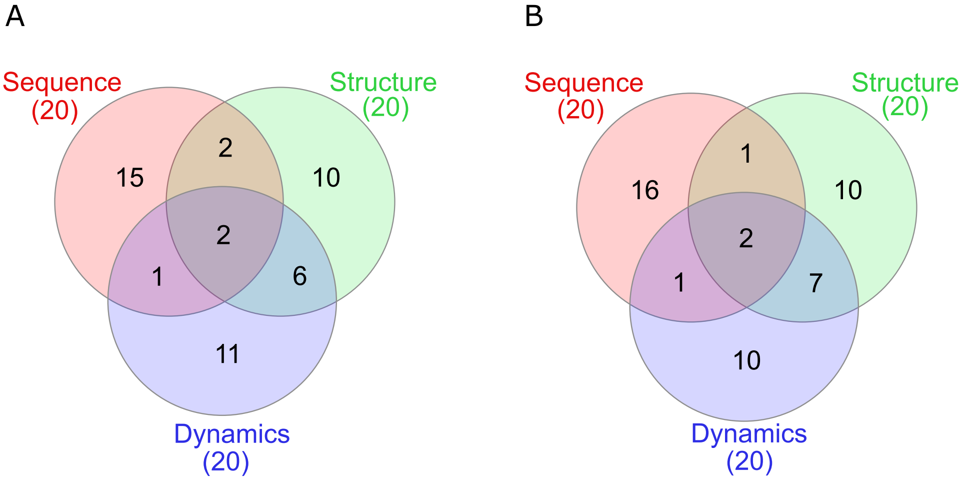

Figure 4.3: Venn diagram showing the overlap of top 20 nodes selected based on the node strength in each of the pair-wise interaction matrices (Fig 4.1) for A. Serine protease and B. DHFR. The selected nodes are given in Tables C.2 and C.3 of C.

4.3.4 Comparison with experimental data

The residues and interactions ranked top in each of the networks were compared with the experimental data available. The residues M42 and Y111 of DHFR which appear in all three approaches were captured in the sector analysis[11] also. M42 being in the adenosine binding subdomain of DHFR can be functionally important. A55 and Y228 are the residues of serine protease that appear in the top 20 residues selected based on each of the sequence, structure and dynamics. A55 and Y228 are related to catalysis and belong to red and green sector respectively. These sectors were defined the study of Halabi et al.[13]. Of the reported catalytic residues of DHFR, I5, M20, D27, K28, F31, L54 and I94, only I5 (Dynamics) and I94 (Structure) are selected at least by one of these methods. In the case of serine protease the catalytic residues D102, H57 and S195 are the ones that appear in the top 20 of one of the methods. While dynamics and structure suggest residues in the core, sequence analysis detect residues that are solvent exposed also. On analysing the importance of residues that were chosen by each method, it can be seen that the blue sector residues which are related to stability, do come in the top 20 of sequence and dynamics, but none in the structure. Also interestingly most of the residues selected by structure are related to catalysis and belong to the green sector. Top 20 residues selected based on SCA has representation from catalytic residues, binding pocket and also the ones related to stability. For DHFR both residues in the binding pocket and the ones related to catalysis appear in the top 20 residues of sequence, structure ad dynamics.

4.3.5 Interaction networks

Interaction networks were created by selecting top 100 residue pairs separated by at east two residues in between. For the commute-time matrix and inter-residue fluctuation matrix, the pairs with lower values were chosen. The connections identified from all three matrices are shown in the network representation in Figure 4.5. The edges common between different networks such as sequence and structure are shown in different colors. While there was a large overlap of edges between structure and dynamics networks, the sequence network had fewer connections common with the structure and dynamics networks. Most of the interacting residue pairs identified based on structure and dynamics were structurally closer compared to the ones chosen based on SC.

4.3.6 Effect of data size

To understand if data size limitations can be captured within the scope of our calculations, we repeated the analyses using trajectory averaged structure,50 ns of MD and selecting sequences with higher sequence identity. The overlap did not change as shown in Figures

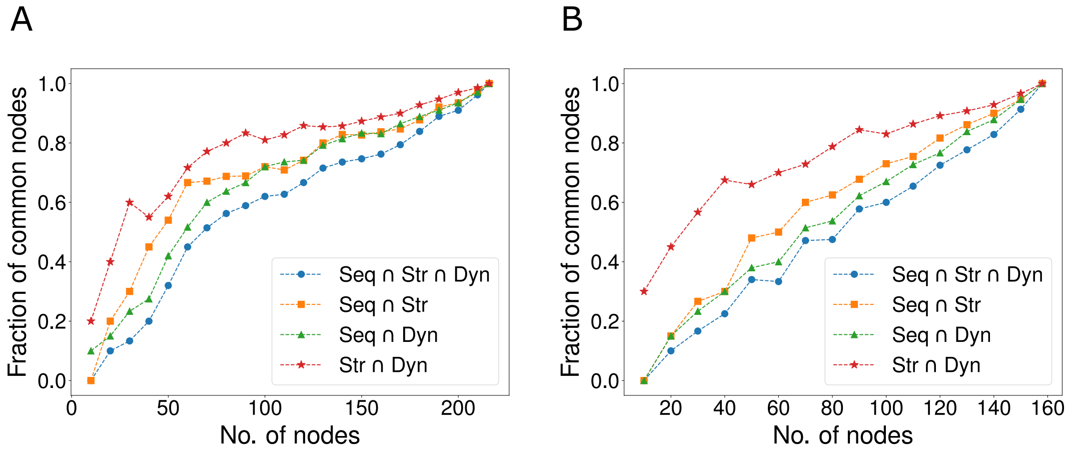

Figure 4.4: The number of top nodes chosen from the networks based on sequence, structure and dynamics is varied and the intersection of these sets of nodes is plotted for A. Serine protease and B. DHFR.

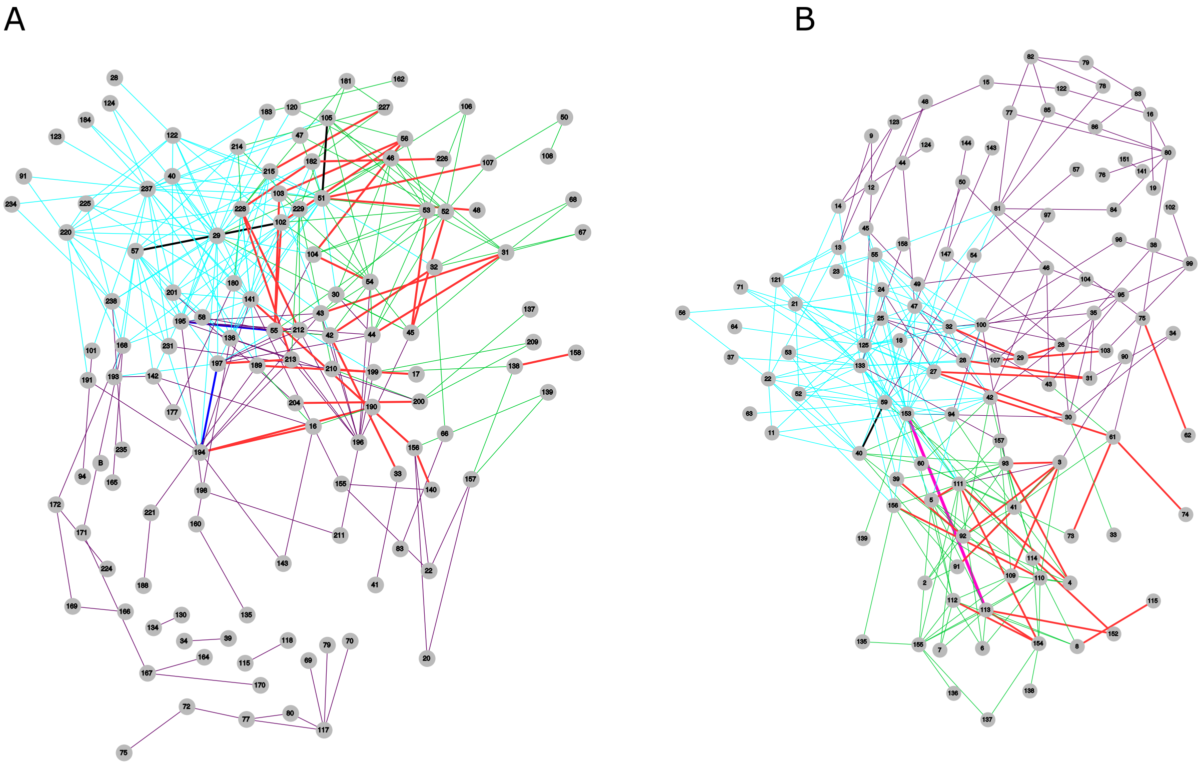

Figure 4.5: Based on each of the pair-wise interaction measures shown in Fig 4.1, amino acid pairs that are separated at least by two amino acids were rank ordered and the strongest 100 connections were selected to construct interaction networks. The union of all three networks is shown for A. Serine protease and B. DHFR. Each edge is weighted proportional to the number of networks it belongs to. Connections in each network or intersection of networks are colored differently.

C.1-C.3 of AppendixC.

4.3.7 Comparison of interaction networks

As the analysis of identifying common edges between the networks when 100 top connections were considered showed poor overlap, we explored whether the networks converge by choosing higher number of connections. In Figure 4.6 the ratio of number of connections that are common between all three networks with the number of connections chosen from each of the networks is shown.

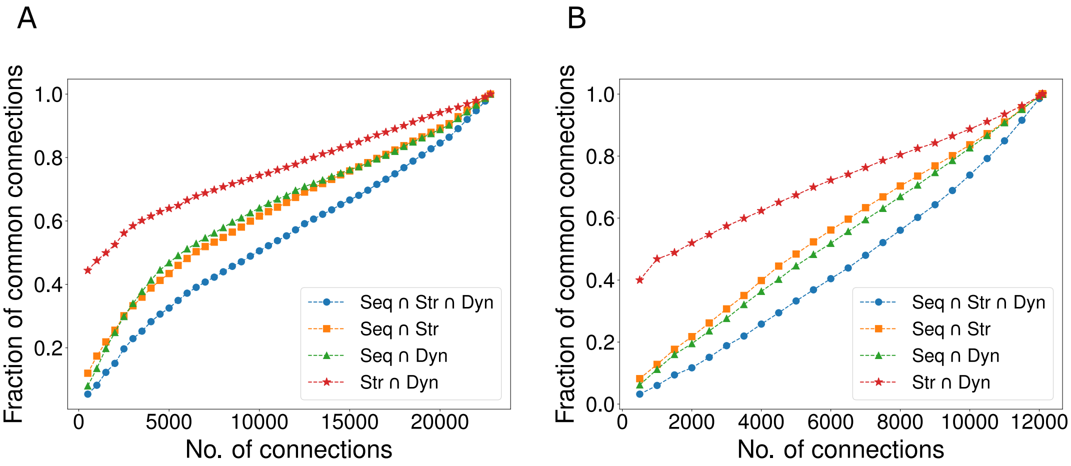

Figure 4.6: Variation in the fraction of common connections in all three networks with the number of connections chosen from each of the networks is shown for A. Serine protease and B. DHFR.

4.4 Discussion

In this work, we followed some of the different methods that are being used to identify the important amino acids in a protein and compared the results. In principle, sequence contains the complete information of how a protein should function. However, the implicit assumption in such a statement is that one is aware of the complete atomic level details of the sequence in the primary structure, rather than just the simplified representation with the 20 alphabets from A to Y. In the absence of such atomic details, the simplified alphabetic representation of the sequences can be complemented with multiple sequences or structural and dynamical information. Thus using multiple sequences, structure and dynamics of the protein one should be able to identify important amino acids. Guided by this fundamental belief, many attempts have been made to identify critical or allosteric interactions. In the spirit of the recent study that compared sequence with dynamics[19], we used different methods based on sequence, structure, or dynamics that are being used for identifying residues which can lead to deleterious effects upon mutation. However, as the results in Fig. 4.4 show, this is not true. The edges that connect pairs of amino acids, possible interpretation being local or allosteric effects, as seen in Figure 4.6, have not much in common. The same is true of the important amino acids we identified. Out of a total of 20 amino acids each that we rank ordered using the three criterion, only 2 of them were common. We curated the critical amino acid mutations that are reported in the literature and all these methods capture a fraction of these amino acid mutations. These differences raise a few questions on the convergence with data size, convergence with the number of predictions made, true positives and predictability.

Convergence with data size. In reality, none of the three methods are complete – basing analysis on one single structure, or molecular dynamics that has inherent sampling limitations or multiple sequences which may not be sufficient in number or in divergence can cause limitations to the results that can be obtained from these different approaches. It is not easy to increase the data size to a point where these limitations will not exist. To check if the data size, within the limits of the calculations we could perform easily, affects the results, the analyses were repeated with half-the length of the simulation trajectory, half the number of sequences in alignment, and a different averaged structure obtained from the simulations. Within this range of variation, there were no (or no significant) differences in the amino acids or edges that we could find, which leaves the question whether the discrepancies can be solved by a reasonable extension of the data size.

Convergence with size of selection. The 20 amino acids chosen from each of the methods was an illustration, although there was no specific meaning associated with the number 20. By varying the number of amino acids that are chosen, as one may expect, the overlap between the three predictions continues to increase, trivially becoming perfect when all the amino acids are chosen. However, surprisingly, even when top 50% of the amino acids were chosen from each of the overlap still does not cross 70% Since the number of edge connections scale as N^2, where N is the number of the amino acids, convergence of the edges when a certain number of selections are made is slower compared to the convergence of the amino acids.

False positives. Because of the differences in the methods, it thus becomes imperative that while inferring the important amino acids or edges from calculations, one must be aware of the number of selections that should be made before the experimentally observed interesting amino acid or pair-wise correlation appears in the analyses. Otherwise, there is a risk that one may be ignoring several of strongly correlated false positives before identifying a medium correlation that is experimentally observed. Alternatively said, the predictive capacity of the correlations rather than the a posteriori justification needs to be evaluated.

Complementary predictions. Until the differences in the predictions from the approaches can be resolved, another way to consider these different predictions is to treat them as being complementary as many approaches now are using sequence, structure, and dynamics based information in the models driven by artificial intelligence to make predictions of the mutational effects.

4.5 Conclusion

We performed a detailed comparison of the results from structure, sequence and dynamics approaches to identify important residues and interactions. It is found that the nodes, interactions chosen according to each of the aspects do have overlap, but many of them are unique to each. Though this fact appears to be trivial, there was no systematic analysis and quantification of the same before. Interestingly, the selected nodes or edges do not converge completely even when 75% of them are selected. The convergence behaviour is sublinear though much higher than a random selection would have. Our study suggests the importance of considering all three aspects to understand the role of an amino acid in the function of a protein.

Bibliography

[1] W. A. Eaton, V. Munoz, S. J. Hagen, G. S. Jas, L. J. Lapidus, E. R. Henry, and J. Hofrichter, "Fast kinetics and mechanisms in protein folding," *Annual review of biophysics and biomolecular structure*, vol. 29, no. 1, pp. 327–359, 2000.

[2] D. B. Ritchie and M. T. Woodside, "Probing the structural dynamics of proteins and nucleic acids with optical tweezers," *Current opinion in structural biology*, vol. 34, pp. 43–51, 2015.

[3] E. Shakhnovich, "Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet," *Chemical reviews*, vol. 106, no. 5, pp. 1559–1588, 2006.

[4] J. D. Bryngelson and P. G. Wolynes, "Spin glasses and the statistical mechanics of protein folding," *Proceedings of the National Academy of Sciences*, vol. 84, no. 21, pp. 7524–7528, 1987.

[5] C. D. Snow, H. Nguyen, V. S. Pande, and M. Gruebele, "Absolute comparison of simulated and experimental protein-folding dynamics," *nature*, vol. 420, no. 6911, p. 102, 2002.

[6] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood,

J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, *et al.*, “Atomic-level characterization of the structural dynamics of proteins,” *Science*, vol. 330, no. 6002, pp. 341–346, 2010.

[7] D. M. Fowler and S. Fields, “Deep mutational scanning: a new style of protein science,” *Nature Methods*, vol. 11, pp. 801–807, AUG 2014.

[8] J. R. Schnell, H. J. Dyson, and P. E. Wright, “Structure, dynamics, and catalytic function of dihydrofolate reductase,” *Annu. Rev. Biophys. Biomol. Struct.*, vol. 33, pp. 119–140, 2004.

[9] J. A. Capra and M. Singh, “Predicting functionally important residues from sequence conservation,” *BIOINFORMATICS*, vol. 23, pp. 1875–1882, AUG 1 2007.

[10] S. W. Lockless and R. Ranganathan, “Evolutionarily conserved pathways of energetic connectivity in protein families,” *Science*, vol. 286, no. 5438, pp. 295–299, 1999.

[11] K. A. Reynolds, R. N. McLaughlin, and R. Ranganathan, “Hot spots for allosteric regulation on protein surfaces,” *Cell*, vol. 147, no. 7, pp. 1564–1575, 2011.

[12] D. Pincus, O. Resnekov, and K. A. Reynolds, “An evolution-based strategy for engineering allosteric regulation,” *Physical biology*, vol. 14, no. 2, p. 025002, 2017.

[13] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, “Protein sectors: evolutionary units of three-dimensional structure,” *Cell*, vol. 138, no. 4, pp. 774–786, 2009.

[14] C. Chennubhotla and I. Bahar, “Signal propagation in proteins and relation to equilibrium fluctuations,” *PLOS Computational Biology*, vol. 3, pp. 1716–1726, SEP 2007.

[15] J. L. Radkiewicz and C. L. Brooks, “Protein dynamics in enzymatic catalysis: exploration of dihydrofolate reductase,” *Journal of the American Chemical Society*, vol. 122, no. 2, pp. 225–231, 2000.

[16] T. Ichiye and M. Karplus, “Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations,” *Proteins: Structure, Function, and Bioinformatics*, vol. 11, no. 3, pp. 205–217, 1991.

[17] O. F. Lange and H. Grubmüller, “Generalized correlation for biomolecular dynamics,” *Proteins: Structure, Function, and Bioinformatics*, vol. 62, no. 4, pp. 1053–1061, 2006.

[18] H. Kamberaj and A. van der Vaart, "Extracting the causality of correlated motions from molecular dynamics simulations," *Biophysical journal*, vol. 97, no. 6, pp. 1747–1755, 2009.

[19] B. Lakhani, K. M. Thayer, E. Black, and D. L. Beveridge, "Spectral analysis of molecular dynamics simulations on pdz: Md sectors," *Journal of Biomolecular Structure and Dynamics*, pp. 1–10, 2019.

[20] P. Calligari, M. Gerolin, D. Abergel, and A. Polimeno, "Decomposition of proteins into dynamic units from atomic cross-correlation functions," *Journal of chemical theory and computation*, vol. 13, no. 1, pp. 309–319, 2016.

[21] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, *et al.*, "The pfam protein families database in 2019," *Nucleic acids research*, vol. 47, no. D1, pp. D427–D432, 2018.

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

PART II

Mutational Effects Prediction and Interpretation

Chapter 5

Deep2Full: Evaluating Strategies for Selecting the Minimal Mutational Experiments for Optimal Computational Predictions of Deep Mutational Scan Outcomes

Abstract

Performing a complete deep mutational scan with all single point mutations may not be practical, and may not even be required, especially if predictive computational models can be developed. Computational models are however naive to cellular response in the myriads of assay-conditions. In a realistic paradigm of assay context-aware predictive hybrid models that combine minimal experimental data from deep mutational scans with structure, sequence information and computational models, we define and evaluate different strategies for choosing this minimal set. We evaluated the trivial strategy of a systematic reduction in the number of mutational studies from 85% to 15%, along with several others about the choice of the types of mutations such as random versus site-directed with the same 15% data completeness. Interestingly, the predictive capabilities by training on a random set of mutations and using a systematic substitution of all amino acids to alanine, asparagine and histidine (ANH) were comparable. Another strategy we explored, augmenting the training data with measurements of the same mutants at multiple assay conditions, did not improve the prediction quality. For the six proteins we analyzed, the bin-wise error in prediction is optimal when 50-100 mutations per bin are

used in training the computational model, suggesting that good prediction quality may be achieved with a library of 500-1000 mutations.

5.1 Introduction

Mutations are changes in the nucleotide sequence of an organism, and its effects may be noticeable across the scales from protein expression, cellular or organismal level. Most mutations are usually found to be neutral or deleterious across the scales, while a very few of them turn out to be beneficial i.e. confer an increase in phenotypic fitness.

While mutations at the active sites of enzymes are relatively easy to interpret, understanding how distal mutations affect the catalytic activity is a challenge on its own. Predicting a change in cellular or organismal fitness upon a single mutation in proteins is further complicated, since fitness is a downstream effect and an immediate correlation with changes in structural stability and dynamics of the protein may not be easy. However, such an understanding will have an enormous impact, whether it is for identifying disease causing mutations in the human genome or for designing antibiotics. The development of high-throughput technologies has driven newer and massively parallel approaches in the exploration of mutational landscapes at a cellular phenotypic level. Methods such as deep mutational scan[1, 2] or site saturation mutagenesis[3] now made it possible to study the fitness consequences of a very large number ($\sim 10^5$) of independent mutations of the same protein.

But performing such experiments is highly resource demanding. Further, as the interest in the study of simultaneous multiple mutations increases, such as in the case of drug resistance and compensatory mutations,[4] the number of mutational studies required will increase by orders of magnitude. Thus, alternative or complementary approaches that quantify the fitness effects of a wide range of amino acid mutations have to be developed.

Several computational tools have been developed to predict the functional effects of mutations: SIFT[5] is based on evolutionary information obtained from the sequences of proteins and their homologs whereas SNAP2[6], PON-P2[7] use other features such as functional annotations along with evolutionary information. Tools such as SNPs3d[8] and Polyphen[9] use information about the 3D structure of the protein also. Condel[10], CADD[11], REVEL[12] and PON-P[13] are predictors that combine the predictions of other tools. Unsupervised methods using sequence covariation (EVmutation)[14] proposed statistical energy scores to be correlated to the fitness effects of mutations, and newer developments in this methodology (DeepSequence)[15] exploit the latent variables to improve the predictions. Another Global epistatic model GEMME has been shown to have better quality of prediction for viral proteins.[16] Recently, deep mutational scan data from different proteins was used for developing a global quantitative model for mutational effects predictions (Envision),[17] which was then used for predicting the effects of all possible single amino acid substitutions in the proteomes of human, mouse, frog, zebrafish, fruit fly, worm, and yeast.

None of these models has the flexibility to adapt when the assay conditions are changed. There have been Proteins Specific Predictors (PSP)[18] which are developed by training on data of specific proteins to classify mutations. Making quantitative predictions of the downstream effects of mutation under an external selection pressure is not easy. While it may be too soon for computational methods to completely replace wet-lab experiments, they are certainly at a stage where they can be used to reduce the number of experiments required and hence the costs of generating such large data sets. The next paradigm in the evolution of the models is thus a combination of partial data from deep mutational scans with computational models. Recently it was demonstrated that the large fractions of data missing from mutational scans can be imputed[19, 20] using machine learning approaches. It is thus clear that exploiting the information about the system and the mutations, one can predict the effects of missing mutations. Continuing on a similar theme, we explore a complementary question about better ways of designing DEEP mutational scan to develop predictions for a FULL mutational scan (DEEP2FULL). Specifically we ask if there is a better strategy to design the experiments with minimal number of mutations and to prioritize experiments rationally, and yet achieve the best possible predictions for the rest of the mutations. In this work, we use publicly available deep mutational scan data on six proteins and illustrate the outcomes of a few strategies we define for choosing the minimal set of mutations.

5.2 Results

5.2.1 Neural network models for predicting fitness

To computationally predict the outcomes of deep mutational scans we developed artificial neural network (ANN) models, using variables which can describe the physico-chemical properties of the wild type amino acid and the substitutions, and partial experimental data on the fitness consequences of the mutations. Seventeen different descriptive parameters (**Methods** section), including 4 parameters derived from the protein structural information, 7 variables from sequence information and 6 others from co-evolutionary information were used in our models. The experimental data we used consisted of relative fitness of the mutant cells with respect to the wild type under selection pressure from different stressors or their concentrations. ANN is similar in philosophy to the goal of predicting the downstream effects of the mutations, as the clarity of what happens at the intermediate stages, also known as layers, is compromised in favor of the end results it generates. While it lacks the simplicity of a linear regression model, it can in principle embody all the complex non-linear interactions that occur at the different stages of the effect propagation, starting from the mutation and ending with the change in fitness. Although several machine learning approaches such as random forests[19] may be useful for making predictions, we chose to work with artificial neural networks. We used feedforward neural network with Levenberg-Marquardt back-propagation algorithm implemented in the Neural Network Toolbox of Matlab along with the early stopping criterion for termination of training. For each data set chosen for modeling, neural network models were built by subdividing it into training, validation and test sets. Apart from the input and output layers, all neural networks had a single hidden layer and the number of neurons in this layer was chosen based on the coefficient of determination (R^2) for the training and validation set predictions (**Methods**).

5.2.2 Impact of sampling size on the model's predictive ability

The first strategy we evaluated was a systematic reduction of the size of the experimental data that was used to train the model. From the complete mutational scan data that was available, a set of randomly chosen variants was used for training and validation and a systematic reduction in the size of this set (85%, 50%, 25% and 15%, respectively) was made for developing four different models. We analyzed six proteins - β-lactamase,[21] aminoglycoside 3'-phosphotransferase (APH(3′)-II),[22] heat shock protein 90 (Hsp90),[23] mitogen-activated protein kinase 1 (MAPK1),[24] ubiquitin-conjugating enzyme E2 I (UBE2I)[19] and thiamin pyrophosphokinase (TPK1).[19] The selection was based on the criterion that data on at least 2500 mutations are available although the assays measured different phenotypes. Some of these experiments measured a change in the average rate of cell division upon mutation,[19] while others measured a consequent variation in the population.[21] The predictions of the model developed using 85% of data for β-lactamase is shown in Figure D.1 of Appendix D and a comparison of the same with experimental data in Figure D.2 of Appendix D. Results from models trained on smaller data sets are summarized in Figure 5.1 for β-lactamase and for the other five proteins in Figures D.3 to D.5 of Appendix D. As expected, the overall quality of predictions improves with increase in the data used for training although the improvement is sublinear (Table D.1 of Appendix D). As can be seen from these results, except for the case of TPK1 the Pearson correlation between predicted and experimental fitness begins to saturate when more than 50% of the data is used for training the model.

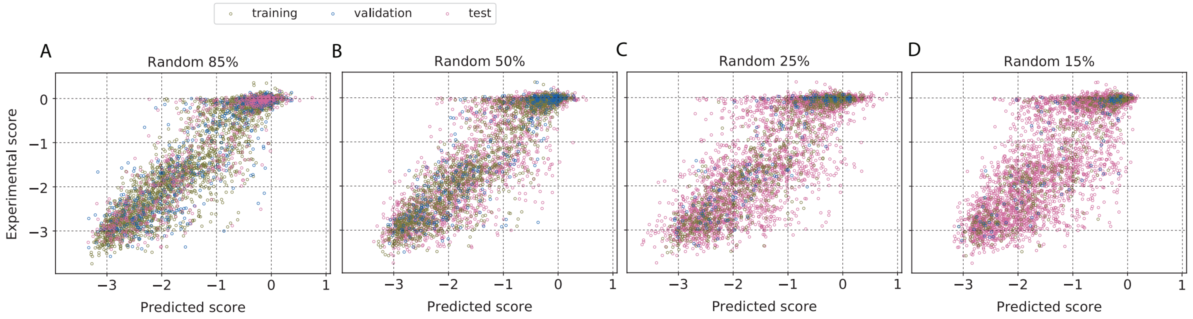

Figure 5.1: Systematic increase of training data size improves prediction quality. The experimental data on the relative fitness of *E. coli* with mutations in β-lactamase was modeled. The fraction of the complete data that was used for training and validation was systematically reduced in four steps from 85% to 15% to see how the quality of computational predictions of fitness changes. It can be seen that the quality of predictions when trained with 50% is comparable with the one trained at 85% data. The prediction quality is tabulated in Table D.1 of Appendix D. Results from predictions of other proteins are in Figures D.3 to D.5 of Appendix D.

5.2.3 Comparing complementary ANH and four other mutational scans trained on 15% of data

One common feature of the four models developed above is that they are all trained on mutations randomly selected from across the sites and possible substitutions (random scans). We further explored if using systematically chosen mutations in model development can help improve the prediction quality. We performed these analyses with the smallest amount of data to have a better chance of observing the differences. We first used the fitness scores from alanine scan, and predicted the outcomes for all other 19 mutational scans. However, in our search for the minimal and predictive data set, alanine scan data was not satisfactory ($R^2_{test} = 0.38$, Figure D.6 of Appendix D). Hence we performed other comparative analyses starting with an augmentation of the alanine scan. It was recently discovered[25] that in multiple deep mutational scan data sets the fitness changes upon mutation to any amino acid is best correlated statistically with the fitness scores associated with asparagine (N) and histidine (H) substitutions. Taking a cue from this observation, we combined the commonly used alanine (A) scan, with asparagine (N) and histidine (H) scans, thus choosing one from each charge type - hydrophobic (A), polar (N) and charged (H), to develop an ANH scan. We then used ANH mutational scan data which is 3/20 or 15% of the full mutational scan data, as a strategy to train the neural network model and to predict the remaining 17 amino acid scan results at every site. The ANH scan data was further divided as 85% for training and 15% for validation of the model. As seen in Figure 5.2**A** and Table D.2 of Appendix D, the fitness predictions

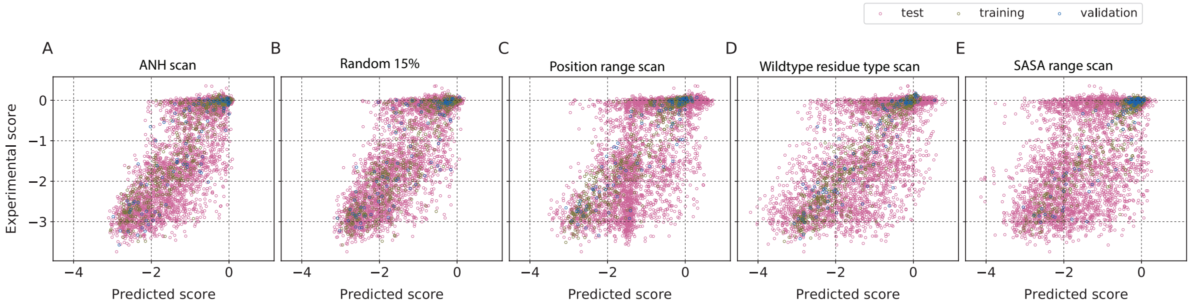

Figure 5.2: Representation of the types of mutations in the training set influences the results. A comparison of the different strategies we used for choosing the training set with 15% data completeness. In an extension of the concept of alanine-scan, the fitness outcomes from alanine (A), asparagine (N) and histidine (H)-scans at each amino acid position were used as the training set, and the fitness scores for all other 17 mutations at every site were predicted. The results were compared to other strategies that used random (Random 15%) or site-directed protocols (position range scan, wild type residue type scan and SASA range scan) for choosing the minimal set required for training. The results suggest that choosing mutations randomly or performing an ANH scan is better than scanning all mutations at a few positions.

improve relative to the one obtained by training on alanine scan data ($R^2_{test} = 0.62$). The results from training the models with either ANH-scan or a random scan, both with 15% data, are comparable, with one working slightly better than the other depending on the protein.

We explored a few other systematic mutagenesis schemes based on the concept of site-directed mutagenesis. We asked if having the data for all 19 mutations at a few positions could improve the prediction quality. We used three different ways of identifying these positions: 1) *Position range scan* - Residue positions were randomly chosen to have an approximately uniform sampling of the sites along the primary sequence; 2) *Wild type residue type scan* - Depending upon the distribution of wild type amino acids, in this scan wild type positions were chosen to ensure that there is a nearly uniform representation of the 20 amino acids in the training set; 3) *SASA range scan* - Residue positions were chosen in such a way that the distribution of solvent accessibility is uniform over the training and validation sets. The idea was to have representation from the residues with different levels of solvent exposure in the training set. The results for β-lactamase are shown in Figure 5.2 and those for the other five proteins are summarized in Figures D.7 to D.9 of Appendix D. The results for all three position based scans, all trained on 15% data, were poorer than those from a random or ANH scan.

5.2.4 Augmenting data with transverse assay conditions

We also investigated whether with the same number of mutations, the prediction quality could be improved by using transverse data from different assay conditions. The rationale for evaluating this strategy was to compensate for the number of mutants with the number of cultures with different stressor concentrations. We trained our models using 15% of the mutational data, but with the fitness changes measured at six different drug concentrations, [21] thus enhancing the total data used for training by 6-fold. Plotting the fitness change for each mutation with *log [ampicillin]* displayed a regular sigmoidal pattern in the dose-response curve, thus raising the possibility that the augmentation brings more structured data and improves predictability. We compared the mutational effects predictions for the studies at 2500 μg/ml using two models, one trained on data from six different concentrations and the other trained only on the data from experiments performed at concentration 2500 μg/ml. However, contrary to our expectations as shown in Figure 5.3, there was no significant difference in the prediction quality by using data at different concentrations for training. The same was also true for the predictions of the mutational effects at 650 μg/ml drug concentration.

5.2.5 Variable importance and models with fewer variables

The primary aim of the work was to reduce the experimental data needed for building the model. However, conceptually it is also interesting to ask if the role of different predictive variables used in the model can be quantified and if the model itself can be simplified. We illustrate the relative importance of the different descriptive variables using our calculations on the fitness (dis)advantage in *E. coli* exposed to ampicillin, conferred by the single point mutations in TEM-1 β-lactamase,[21] although the scope of the analysis is general. For investigating the contribution of individual input variables in the predictions, the input variable was kept fixed at its mean value for all the samples and the network was retrained. The change in mean squared error (MSE) on the removal of a variable is used for quantifying the importance of that input variable. Figure 5.4 shows the difference in MSE when each variable is replaced by its mean value. BLOSUM which represents the substitution effects based on evolutionary data has the highest contribution to the predictions. Hydrophobicity index of the amino acid to which the mutation is made and the average commute time are the other variables with significantly higher contributions. In addition to the 17 variables, we also added the statistical coupling energy[14] as an additional variable to see if it improved the correlation between the predictions and the observations. No improvement was noticed, possibly because other variables including the ones from co-evolution data already implicitly accounted for this factor (Table D.3 of Appendix D). Since the proximity of an amino acid to the

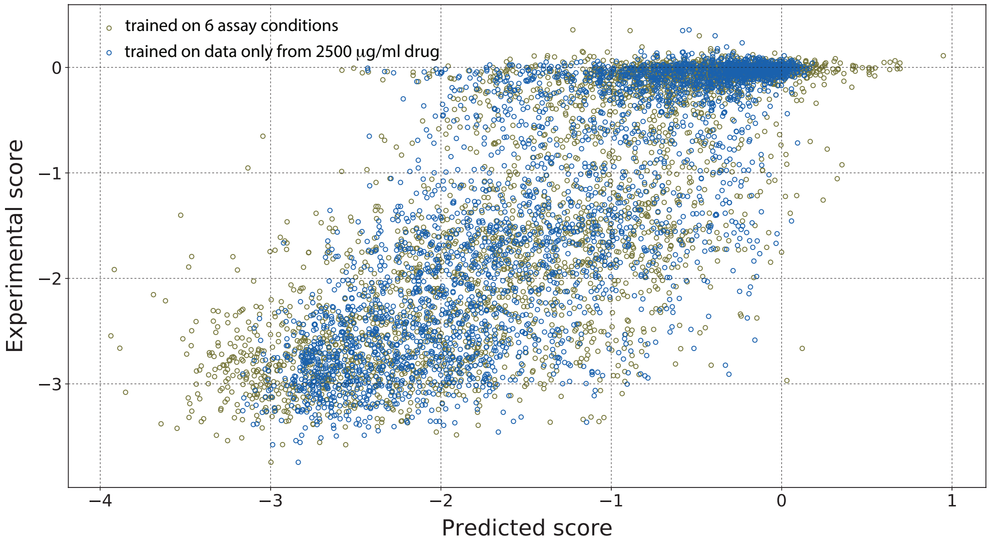

Figure 5.3: Augmenting with scores at different assay conditions did not improve predictions. At 15% mutational completeness, the data size was augmented by combining data from six different assay conditions. There was no improvement in the prediction quality although the data was enhanced 6-fold. The R^2_{test} with and without data augmentation was 0.61 and 0.66 respectively. More detailed results are in Figure D.10 of Appendix D. Similar analysis was performed by developing model trained on data from the 650 µg/ml drug concentration assay. In this case also the predictive ability of the models trained at only one concentration or at multiple concentrations was similar.

catalytic site could be of high functional significance, we developed a model with this factor as an additional descriptive variable. The catalytic residues were identified in β-lactamase and the distance of every amino acid to the nearest catalytic residue was computed. This additional input variable did not improve the predictions either. As in the case of statistical coupling energy, the information contained in this variable could be represented by other variables like conservation, number of contacts and commute time. So statistical coupling energy and distance from catalytic sites were not used in any other analysis in this work. We also analyzed the contributions at a coarse level, creating neural network model for alanine scan mutations using only (1) sequence based variables and (2) structure based variables. The sequence based model performed better than the structure based one, R^2 values being 0.54 and 0.25 respectively for the sequence and structure based models for the test set chosen from the alanine scan data set.

We selected fewer variables and developed minimal models using two different measures to rank the individual variables: Pearson correlation of the individual variables with the measured fitness and the change in MSE on replacing the variables with their averages.

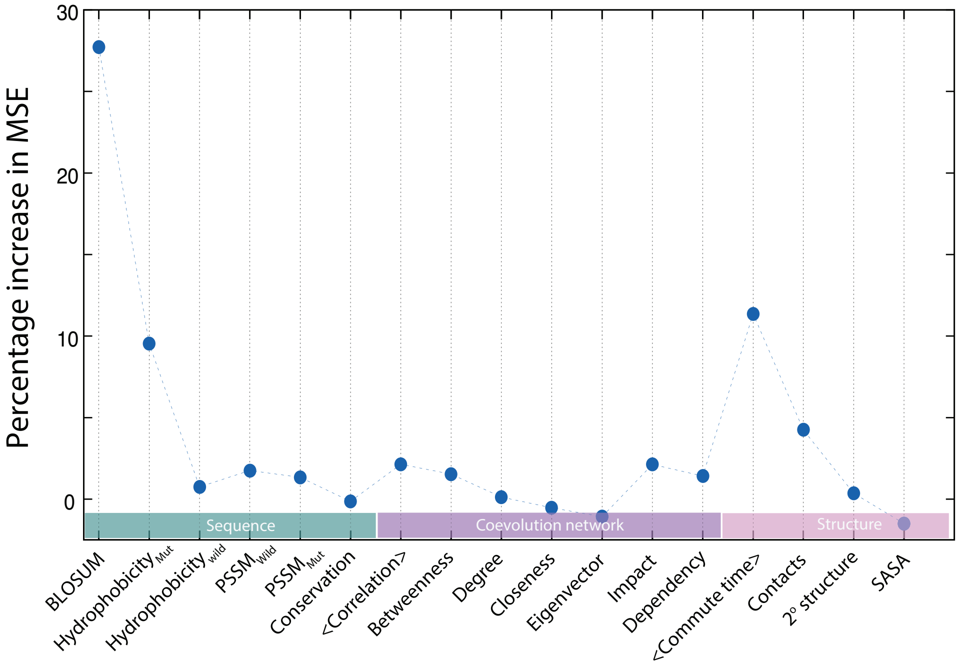

Figure 5.4: A few variables contribute significantly. The relative importance of different variables in the predictive model trained with 85% data from β-lactamase mutations was evaluated. The sensitivity of the model to a variable was quantified as the percentage increase in the mean squared error (MSE) between the prediction and the experimental values when the variable was replaced with its average calculated across all mutations. BLOSUM score, average commute time and hydrophobicity of the mutant have the highest contribution while some of the variables have little contributions in the model. None of the variables we used is perfectly correlated to any other variable, however, the poor contributions suggest that they could be correlated to a non-linear combination of other variables.

Using these two criteria models were developed using 7 and 6 variables respectively (**Methods** section). Average correlation, average commute time, number of contacts of the wild type amino acid, and BLOSUM score for the substitution were the most relevant variables according to both of these criteria. The results obtained (Figure D.11 of Appendix D) from these two reduced models are of comparable quality to the ones constructed with 17 variables. However, in the interest of the scope of the present work which is about reducing data rather than reducing variables all our analyses are presented with the results of model trained on 17 predictive parameters.

5.2.6 Quality analysis of output and input

The quality of predictions in our analysis was verified based on three different measures - (1) the overall $R^2_{prediction}$, Root Mean Square Deviation (RMSD) and Pearson correlation, all three metrics suggested that the quality of our predictions were comparable with other models which use partial data for prediction (Table D.5 of Appendix D). It is notable that R^2, which is very sensitive to outliers also had shown that the predictions are reasonable even at low data completeness, (2) the prediction data was segregated either based on the amino acid before or after mutation. The outcomes for some of the amino acids are relatively poor as seen from the individual regression plots (Figures D.12 and D.13 of Appendix D). Amino acid wise prediction quality can be summarized using their Pearson correlation values also as shown in Figure 5.5. We further analyzed and found that the quality of predictions for different amino acids (Figure 5.5B) was not correlated with their frequency in the training set. It can be seen that the effects of some amino acid mutations do not span the entire range of fitness scores, hence predictions could not be improved. (3) For a predicted fitness, the variation in the experimental values. This is summarized in Figure D.2 of Appendix D with histograms of experimental fitness generated from the predicted fitness variation around -3, -2, -1 and 0. While these histograms show a variation relative to the predicted fitness, it must be noted that even in different trials of the experiment, there is a significant variation. We also investigated the prediction quality for amino acids with different solvent exposure (Figure D.14 of Appendix D). As can be seen, in general the predictions were better in quality for the solvent exposed residues. This could be because of the lower variability in fitness scores at higher SASA range.

Across the six proteins we studied, the quality of the predictions varied. We checked if it is possible to define a measure for the quality of input data which forces a requirement on the size of the data set used for training. It is apparent from the experimental data that the range of measured fitness varies depending on the protein and stressor concentration. It also appears from our results that the prediction quality may be slightly better when the fitness effects in the experiments span a broader range. In an attempt to clarify these effects of input data quality and size, we defined a quality metric of the input data as the ratio of the range over which the training data spans and the standard deviation of the data centered around what appears to be the neutral mutations. The motivation for choosing such a metric is that the wider range of mutational scores and the separability of neutral mutations from the rest will lead to improved predictions. Mutational effect scores for β-lactamase measured under different concentrations of ampicillin (2500 μg/ml , 625 μg/ml, 156 μg/ml, 39 μg/ml) were available in the study by Stiffler *et al.*.[21] We developed models for each of these data sets and the quality of inputs and output is plotted in Figure D.15A of Appendix D showing a correlation that, as the quality of the input data increases, the prediction quality improves as well. Similar input-output quality analysis was made for all proteins and random scans with systematically increasing data as shown (Figure D.15B of Appendix D). The results although not very conclusive suggest

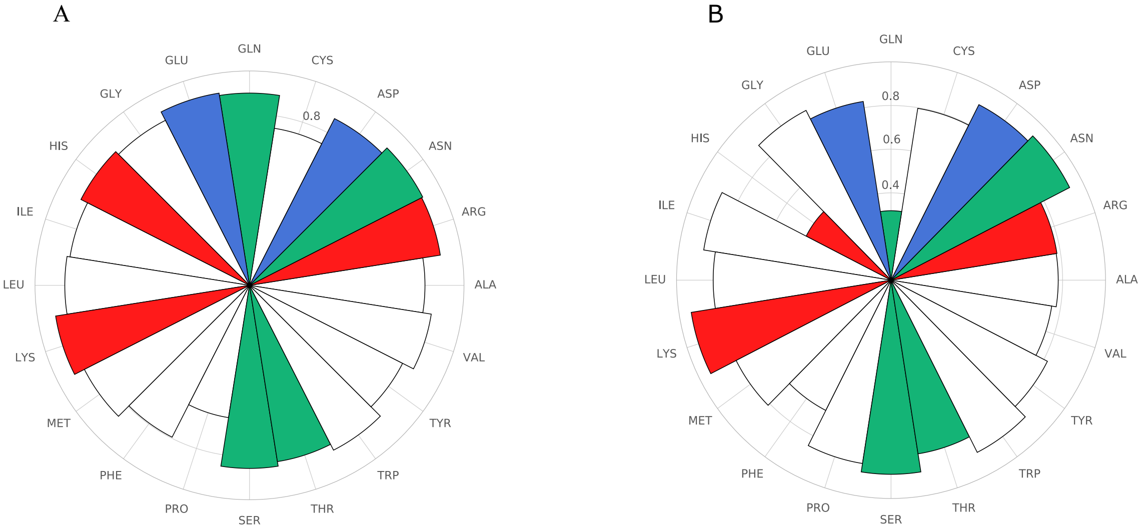

Figure 5.5: Random scan obtains comparable predictions for different amino acids. The test set of random 25% scan was sorted based on the amino acid after mutation and the amino acid in the wild type. The quality of predictions as quantified by Pearson correlation is shown for (A) the amino acid after mutation (B) the amino acid in the wild type. Amino acids are colored according to their type: red (positively charged), blue (negatively charged), green (polar), white (hydrophobic). The random scan results in roughly uniform quality of predictions for all substituted amino acids.

that the predictability of the fitness may be improved by using data obtained at high stressor concentrations.

5.3 Discussion

5.3.1 Scanning strategies: Training with reduced data

The goal of Deep2Full was to evaluate how the data required for developing the computational model can be systematically reduced, and if for a given size of data a systematic sampling can improve the quality of predictions. Having additional data of fitness scores from other assay conditions such as different drug concentrations with the smaller training sets did not improve the prediction quality. Quite intuitively, the quality of prediction for all the proteins improves with increase in the training data from 15% to 85%. We used two different metrics of quality - RMSD and Pearson correlation. Although the prediction quality increases, the improvement does not scale linearly with the data size (Figure D.16 of Appendix D). With both these measures we could see a sign of saturation when more than 50% of the data was used for training. We also analyzed the error across the range of experimentally measured fitness. Ten bins of equal widths were created dividing the experimentally observed fitness in each protein and RMSD for the test set was calculated for each bin. For all the proteins, the error systematically increases as the expected effects of mutations increase (Figure D.17 of Appendix D). To further understand this RMSD, we plotted RMSD for each bin relative to the number of data points in this range that were used for training the model (Figure D.18 of Appendix D). The two observations that come out are a power-law behavior in the error and that increasing training data from each bin beyond 50 to 100 did not improve the predictions significantly. This optimal choice along with a 10-bin division suggests about 500-1000 mutations to be required for developing neural network models. However it appears that one can increase the data used for training in each bin selectively to achieve this optimality. This training data size is approximately 15-25% of the data in the cases we studied.

5.3.2 Scanning strategies: Choosing mutations for training

We asked if there is a better way of choosing the mutations that are used for training the model, guided either by the physico-chemical factors or on the experimental ease of obtaining those mutations. We performed these analyses at the lowest level of data (15%) that yields reasonable predictions, the rationale behind it being that any differences in the strategies will be pronounced and easy to infer. A consistent pattern noted in our study of all six proteins is that at 15% completeness, randomly selecting variants as well as the systematic ANH scan yielded results of comparable quality. The alternative of training the model by selecting all 19 substitutions at a few positions, selected for a representation across SASA or wild type amino acid range were poorer in predictive ability. These trends were consistent in the different metrics we used for determining the quality of predictions - RMSD and Pearson correlation (Table D.2 of Appendix D). It is clear from this analysis that in developing a model, a training set having representation from every position in the protein is much more valuable than having all substitutions at a few sites.

The underlying objective of this exercise was to see if the efforts of constructing the mutant libraries, clones and sequencing can be reduced, without compromising on the quality of the learning. In our model, a random scan implies a random and unbiased choice of the mutation from across the primary sequence where a transition from any amino acid to another is feasible, such as the ones that could be achieved with mutagenesis techniques like POPCode[19] and single-site saturation mutagenesis.[26] We also investigated another scenario where the SNPs may be generated by an error-prone PCR (epPCR), which has an inherent bias against certain mutations.[27, 28] When the model was trained on data set same in size to that of Random 25% scan, but the mutants being chosen from SNPs which were theoretically considered achievable by epPCR,[29] the prediction quality was comparable (Figure D.19 of Appendix D and Table D.1 of Appendix D) for all the six proteins we studied. A detailed cost-benefit analysis considering the number of experimentally realizable mutations which could range from 15% to 85% of data completeness and the accuracy of predictions, will be required to choose between epPCR and site-directed mutagenesis techniques for generating mutant libraries.[30]

5.3.3 Need for hybrid models

Scoring models such as SNAP have been used for classifying the mutations as fitness-neutral or non-neutral.[31] Other recent co-evolution based models have shown a good correlation of the fitness variations observed in deep mutational scans with the predictions of the evolutionary statistical energy[14] and DeepSequence[15]. As shown in Figure D.20 of Appendix D, for the specific example considered, the relations between the scores and cellular fitness using 15% and 50% of randomly selected mutations are at least as noisy or worse than the models we used. Envision[17] was a model that was ambitiously developed to make unsupervised predictions for the deep mutational scan. The model was validated using leaving-one-protein-out protocol. However, as it can be seen from Table D.5 of Appendix D, when using the model for a newer protein such as TPK1 and UBE2I the predictions were not satisfactory. This limitation of the generalized model may be more likely because the mutational effects in proteins are complicated to predict, rather than because of the shortcomings in the specific model. As noted in Table D.5 of Appendix D some of the unsupervised computational predictions that are reported in the literature had good correlations with the experimental data. This raises a question on why the present work focuses on hybrid computational models with minimal experimental data, when the models may possibly be developed with no experimental data. From the deep mutational scan studies on β- lactamase,[21] one can see that the fitness outcomes change with the stressor concentration and with the type of the stressor (stressors: cefotaxime and ampicillin). It was also highlighted in the mutational studies of APH(3′)-II[22] that the fitness landscape depends sensitively on the type of the antibiotic, even when all of them are believed to interact with the same active site. A model that does not use partial experimental data will certainly be insensitive to these differences in the assay conditions. Further, compared to a generic model, one may be able to use highly descriptive protein-specific variables which may improve the predictions.

Newer hybrid models[19] combined partial experimental observations along with other biophysical descriptors to impute the missing data. They could establish that it is possible to achieve predictions that are comparable to the experimental variance across the trials. We used a similar approach of hybrid model and obtained predictions for TPK1 and UBE2I comparable with those from other works[19] (Table D.4 of Appendix D). The predictive ability of our models for the proteins β-lactamase, APH(3′)-II and

Hsp90 are also comparable with that of Envision, where 80% of data was used to develop models for individual proteins[17](Table D.5 of Appendix D). The hybrid models and the strategies will gain prominence as the experimental emphasis shifts towards simultaneous multiple mutations. However, in the present hybrid approach using experimental data and computations one encounters at least two disadvantages that expertise in building computational models and experimental data are required for model development.

5.3.4 Scope and limits of Deep2Full

Although the results of our computational predictions are comparable to those from other hybrid models, the focus of the present work was different. Our goal in this work was not to validate whether the missing data can be predicted, but rather to evaluate if there is a rational way of planning the reduced experimentation. The scope of Deep2Full was to conceive a few ways of designing the minimal number of experiments that will be helpful in training models and evaluating their efficacies in making the predictions for the complete set of variants. It appears that a randomized set of mutations presents the best training set, followed by a charge based training set. While it is preliminary to say, it appears that by optimizing the stressor concentrations in the assays, one may be able to obtain comparable quality of results with fewer mutations. A re-evaluation of the strategies considering the costs or convenience associated with them may be a subject of future work.

Though approaches like Deep2Full which use AI methods can reliably predict the mutational effect, these models are complex and do not help in understanding the relation between inputs and output or help in developing intuitions. In the coming chapters multiple approaches to address this limitation are presented.

5.4 Conclusions

Deep2Full was developed in the context of a new paradigm of hybrid models that train the computational models with partial deep mutational scan data from across assay conditions, to quantitatively predict the fitness outcomes of a full-set of mutations. By combining this phenotypic deep scan data with structure, sequence and co-evolutionary information, the possible outcomes of a full set of deep mutational scan were predicted. We addressed two questions, how much the data size used for training can be reduced and if there is a better way of performing these mutations. To train the models, we found that a representation from all positions of the protein was required. The neural network models which were constructed with seventeen variables from - structure, sequence or co-evolution could in principle be simplified by as few as seven variables, although the model reduction was not the emphasis of the present work. Variation in the experimental data in the different trials, and the choice of the phenotype being measured, such as the differential rate of growth or changes population size, are the limitations the model begins with and this quality of data certain imposes constraints on the size of the data that is necessary for training a reliable model. Regardless, it appears that the best way to enhance the prediction is with a random scan of the sites and substitutions.

5.5 Methods

Data sets chosen: Deep2Full was developed for the deep mutational scan data of 6 proteins: β-lactamase[21], aminoglycoside 3'-phosphotransferase (APH(3′)-II)[22], heat shock protein 90 (Hsp90)[23], mitogen-activated protein kinase 1 (MAPK1)[24], ubiquitin-conjugating enzyme E2 I (UBE2I)[19] and thiamin pyrophosphokinase (TPK1)[19]. Data used for calculations involving mutational effect scores of β-lactamase at different concentrations of ampicillin was obtained from the study of Stiffler *et al.*[21]. Unless mentioned otherwise, the analyses on β-lactamase were preformed on the average of the two trials of the experiments at 2500 μg/ml concentration of ampicillin. For APH(3′)-II, Hsp90 and MAPK1, the computational models were built using the curated data from *Gray et al.*[25] Fitness scores for UBE2I and TPK1 were obtained from the study of *Roth et al.*[19] The total number of mutants used for developing model in each data set was: β-lactamase - 3952, APH(3′)-II - 4234, Hsp90 - 4021, MAPK1 - 4470, UBE2I - 2563 and TPK1 - 3181. For the modeling efforts, we chose to work with proteins for which structural information and data on at least 2500 mutations were available.

Division of data set: The variants for training and validation were chosen according to the strategy in each case. Fitness scores of the chosen variants were then grouped into 3 bins and data points in each bin were divided into training and validation sets in the ratio 85:15. For example, in the random 50% scan, we used 42.5%(= 0.85*50) for training and 7.5%(=0.15*50) for the validation set and the rest for testing. For the random scans the choice of mutations used for developing model is representative of the complete data set, as suggested by the similarity of SASA distributions for the complete data set and the training set which are similar (Figure D.21 of Appendix D).

Choice of parameters: A total of 17 descriptive parameters were used in the our modeling. The *structural variables* for each amino acid that could be calculated from a reference protein structure were included in the model - (1) Solvent accessible surface area (SASA) (2) Secondary structural order, with a binary value 1 if the residue is part of a helix or β-sheet, and 0 otherwise (3) Number of structural contacts an amino acid has with a 4 Å cutoff (4) Average commute time,[32] which reflects the average connectivity of a given amino acid with the rest of the protein. The second group of independent parameters was based on the *sequence information* - (5) BLOSUM substitution matrix

(BLOSUM62) score, which is the probability of substitution of an amino acid by other amino acids inferred from evolutionary information[33] (6) Hydrophobicity on the Kyte-Dolittle scale[34] of the amino acid after mutation (7) Hydrophobicity of the amino acid in the wild type (8) Position specific scoring matrix (PSSM) score for the amino acid after mutation calculated from the multiple sequence alignment (MSA) using PSI-BLAST (9) PSSM score for the wild type amino acid (10) Conservation of the amino acid. The third group of independent parameters was based on the properties of *co-evolutionary networks* that were constructed using the multiple sequence alignment (MSA) of hundreds of homologous proteins. This group is supposed to reflect the importance of an amino acid in an undirected co-evolutionary network - (11) Average co-evolution score of each amino acid (12) Degree centrality, the number of nodes to which a node is connected (13) Betweenness centrality, quantifying the importance of a node in connecting other pairs of residues (14) Closeness centrality, the inverse of the sum of distances to all other nodes (15) Eigenvector centrality, which considers not just the number of connections a node has, but also the connectivity of the immediately connected nodes. Directed network information is also included in the model - (16) Impact factor[35], the number of compensatory mutations required for mutations at a residue position is calculated based on conditional probabilities (17) Dependency factor, which is the counterpart of impact factor is the number of residues which are likely to influence a mutation at a given position. Details about calculation of these parameters are described in the following sub-sections.

Multiple Sequence Alignment(MSA) and inputs calculated using MSA: For β-lactamase MSA was obtained from the Pfam database (Pfam ID: PF13354). Only 208 residues (positions $51 - 260$) of the *E. coli* β-lactamase appeared in the Pfam alignment. So all the calculations and analyses were performed for the fitness effects of substitutions at these positions (3952 data points). For other proteins, homologous sequences were obtained through PSI-BLAST search and were aligned using Clustal Omega.

The variables calculated from MSA are:

Conservation: Conservation was calculated as the percentage occurrence of the most frequently occurring residue at a given position.

Position specific scoring matrix (PSSM): PSSM was calculated from MSA using PSI-BLAST and it quantifies the probability of occurrence of each amino acid at each position of the protein.

Co-evolution network and properties: Multiple Sequence Alignment (MSA) for the protein of interest was truncated to the reference sequence and sequences with a gap frequency less than 20% were used in the analysis. Consensus sequence was generated

using the amino acid with the highest frequency at a given position. Following the Statistical Coupling Analysis protocol, [36] MSA was converted into a boolean sequence, with a 1 if the amino acid is the same as in the consensus sequence and 0 otherwise.
Undirected network: The co-evolutionary relation between two amino acids i and j is calculated as proposed by Halabi *et al.*[36], $C_{ij} = \phi_i\phi_j \; |\langle x_i x_j\rangle_s - \langle x_i\rangle_s\langle x_j\rangle_s|$, where $\phi_i = ln\,((\langle x_i\rangle_s(1-q^{a_i}))\,/\,(q^{a_i}(1-\langle x_i^s\rangle_s)))$, and q^{a_i} is the probability with which the amino acid a_i at position i in the consensus sequence occurs among all proteins. x_i is the i^{th} column in the boolean sequence and $\langle\rangle_s$ denotes the average over sequences.
The co-evolutionary matrix is converted into a network representation using a cutoff c. If $C_{ij} > c$, we consider an undirected co-evolutionary network $i - j$ to be present. In the present analysis weighted co-evolutionary matrix was used and the cut-off chosen was 1. We calculated different centrality measures - eigenvector centrality, degree etc. for the amino acid network described above, using the *igraph* module in python.[37]
Directed network : Using the binary representation of the multiple sequence alignment, we created a directed influence network, in a co-evolutionary sense, with the following conditional probabilities:

$$P(j=1|i=1) = \frac{\text{No. of sequences with } \textit{i=1} \text{ and } \textit{j=1}}{\text{No. of sequences with } \textit{i=1} \text{ and } \textit{j=0} \text{ or } \textit{1}}$$

$$P(j=0|i=0) = \frac{\text{No. of sequences with } \textit{i=0} \text{ and } \textit{j=0}}{\text{No. of sequences with } \textit{i=0} \text{ and } \textit{j=0} \text{ or } \textit{1}}$$

where i and j represent positions. If both $P(j = 1|i = 1)$ and $P(j = 0|i = 0)$ are simultaneously greater than a value P (we used P = 0.8) then position i has an impact on j. A directed network is constructed by identifying all such pairs of residues. In this directed network, the number of outgoing links is considered the impact of an amino acid, and the number of incoming links is considered its dependency. The impact and dependency are supposed to summarize how many simultaneous mutations are forced or forced-upon by a mutation.[35]
Average commute time: The hypothesis that the structural and dynamical connectivity of an amino acid to other amino acids determines the importance of an amino acid has been put forward.[32] The average commute time has been used for identifying hotspot amino acids. The resistance matrix is constructed using the number of atom-atom contacts between amino acids i and j, which are within 4 Å . The resistance matrix is then used for average commute time calculations as per the algorithm suggested in Ref. [32]. All structural variables including average commute time were calculated using the protein structure obtained using the Protein Data Bank (PDB) identifiers: β-lactamase - 1M40, APH(3′)-II - 1ND4, Hsp90 - 2CG9, MAPK1 - 4NIF, UBE2I - 2UYZ and TPK1 -

3S4Y.

Neural network model: All neural network calculations were performed using the Neural Network Toolbox of Matlab2017b. All neural network models had the architecture with an input and output layer and a single hidden layer. The number of neurons in the hidden layer was varied from 2 to 20 for most of the 15% scans, and from 10 to 45 for the other scans where the training sets were larger. Since the initial weights and biases can affect training, for each choice of the number of hidden neurons, 200 neural network models were constructed with random initialization of weights and biases. The predictions from each of these 200 trained models were treated as different trials of the same experiment, and the score for each mutant was calculated as the average of the 200 model predictions. R^2 value for the combined set of training and validation data was monitored with the increase in the number of hidden neurons as illustrated in Figure D.22B of Appendix D. The number of hidden neurons was then chosen as the one with which the R^2 value is the highest (Optimal number of hidden neurons given in Table D.6 of Appendix D). In all these above mentioned calculations we used Levenberg-Marquardt algorithm with mean square error as the performance function for training the network. Early stopping criterion was used to prevent overtraining. The parameters performance goal (*trainParam.goal*), the minimum performance gradient (*trainParam.min_grad*) and maximum number of validation fails before the training is stopped (*trainParam.max_fail*) were set to 10^{-7}, 10^{-8} and 100 respectively. Default values in the *trainlm* algorithm of Neural Network Toolbox of Matlab R2017b was used for all other parameters.

Models with reduced set of variables: The important variables were identified in two ways: (1) Assuming a linear relation between the fitness and input variable, the fraction of variance in the fitness data explained by the input variable is calculated as the square of the Pearson correlation between the input and fitness. Variables with the fraction of variance explained more than 0.1 were chosen to develop the model and were conservation, average correlation, average commute time, contacts, BLOSUM, SASA and PSSM score for the wild type amino acid; (2) Neural network models were developed by fixing each of the inputs to its average value and the percentage increase in the mean squared error upon this is used to quantify variable importance. 6 important variables were chosen based on this: impact, average correlation, average commute time, contacts, BLOSUM and hydrophobicity of the substituted amino acid.

Bibliography

[1] D. M. Fowler, C. L. Araya, S. J. Fleishman, E. H. Kellogg, J. J. Stephany, D. Baker, and S. Fields, “High-resolution mapping of protein sequence-function relationships,” *Nature Methods*, vol. 7, p. 741, SEP 2010.

[2] R. T. Hietpas, J. D. Jensen, and D. N. A. Bolon, “Experimental illumination of a fitness landscape,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, pp. 7896–7901, MAY 10 2011.

[3] L. Zheng, U. Baumann, and J.-L. Reymond, “An efficient one-step site-directed and site-saturation mutagenesis protocol,” *Nucleic Acids Research*, vol. 32, p. e115, 2004.

[4] D. M. Weinreich, N. F. Delaney, M. A. DePristo, and D. L. Hartl, “Darwinian evolution can follow only very few mutational paths to fitter proteins,” *science*, vol. 312, no. 5770, pp. 111–114, 2006.

[5] N.-L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, and P. C. Ng, “SIFT web server: predicting effects of amino acid substitutions on proteins,” *Nucleic Acids Research*, vol. 40, pp. W452–W457, JUL 2012.

[6] M. Hecht, Y. Bromberg, and B. Rost, “Better prediction of functional effects for sequence variants,” *BMC genomics*, vol. 16, no. 8, p. S1, 2015.

[7] A. Niroula, S. Urolagin, and M. Vihinen, “Pon-p2: prediction method for fast and reliable identification of harmful variants,” *PloS one*, vol. 10, no. 2, p. e0117380, 2015.

[8] P. Yue, Z. Li, and J. Moult, “Loss of protein structure stability as a major causative factor in monogenic disease,” *Journal of Molecular Biology*, vol. 353, pp. 459–473, OCT 21 2005.

[9] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, “A method and server for predicting damaging missense mutations,” *Nature Methods*, vol. 7, pp. 248–249, APR 2010.

[10] A. González-Pérez and N. López-Bigas, “Improving the assessment of the outcome of nonsynonymous snvs with a consensus deleteriousness score, condel,” *The American Journal of Human Genetics*, vol. 88, no. 4, pp. 440–449, 2011.

[11] M. Kircher, D. M. Witten, P. Jain, B. J. O’Roak, G. M. Cooper, and J. Shendure, “A general framework for estimating the relative pathogenicity of human genetic variants,” *Nature genetics*, vol. 46, no. 3, p. 310, 2014.

[12] N. M. Ioannidis, J. H. Rothstein, V. Pejaver, S. Middha, S. K. McDonnell, S. Baheti, A. Musolf, Q. Li, E. Holzinger, D. Karyadi, *et al.*, “Revel: an ensemble method for predicting the pathogenicity of rare missense variants,” *The American Journal of Human Genetics*, vol. 99, no. 4, pp. 877–885, 2016.

[13] A. Olatubosun, J. Väliaho, J. Härkönen, J. Thusberg, and M. Vihinen, “Pon-p: Integrated predictor for pathogenicity of missense variants,” *Human mutation*, vol. 33, no. 8, pp. 1166–1174, 2012.

[14] T. A. Hopf, J. B. Ingraham, F. J. Poelwijk, C. P. I. Scharfe, M. Springer, C. Sander, and D. S. Marks, “Mutation effects predicted from sequence co-variation,” *Nature Biotechnology*, vol. 35, pp. 128–135, FEB 2017.

[15] A. J. Riesselman, J. B. Ingraham, and D. S. Marks, “Deep generative models of genetic variation capture the effects of mutations,” *Nature Methods*, vol. 15, pp. 816+, OCT 2018.

[16] E. Laine, Y. Karami, and A. Carbone, “Gemme: a simple and fast global epistatic model predicting mutational effects,” *Molecular biology and evolution*, vol. 36, no. 11, pp. 2604–2619, 2019.

[17] V. E. Gray, R. J. Hause, J. Luebeck, J. Shendure, and D. M. Fowler, “Quantitative missense variant effect prediction using large-scale mutagenesis data,” *Cell systems*, vol. 6, no. 1, pp. 116–124, 2018.

[18] C. Riera, N. Padilla, and X. de la Cruz, “The complementarity between protein-specific and general pathogenicity predictors for amino acid substitutions,” *Human mutation*, vol. 37, no. 10, pp. 1013–1024, 2016.

[19] J. Weile, S. Sun, A. G. Cote, J. Knapp, M. Verby, J. C. Mellor, Y. Wu, C. Pons, C. Wong, N. van Lieshout, F. Yang, M. Tasan, G. Tan, S. Yang, D. M. Fowler, R. Nussbaum, J. D. Bloom, M. Vidal, D. E. Hill, P. Aloy, and F. P. Roth, “A framework for exhaustively mapping functional missense variants,” *Molecular Systems Biology*, vol. 13, DEC 2017.

[20] W. Yingzhou, J. Weile, A. Cote, S. Sun, J. Knapp, M. Verby, and F. P. Roth, “A web application and service for imputing and visualizing missense variant effect maps.,” *Bioinformatics (Oxford, England)*, 2019.

[21] M. A. Stiffler, D. R. Hekstra, and R. Ranganathan, “Evolvability as a Function of Purifying Selection in TEM-1 beta-Lactamase,” *Cell*, vol. 160, pp. 882–892, FEB 26 2015.

[22] A. Melnikov, P. Rogov, L. Wang, A. Gnirke, and T. S. Mikkelsen, "Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes," *Nucleic Acids Research*, vol. 42, no. 14, 2014.

[23] P. Mishra, J. M. Flynn, T. N. Starr, and D. N. A. Bolon, "Systematic Mutant Analyses Elucidate General and Client-Specific Aspects of Hsp90 Function," *Cell Reports*, vol. 15, pp. 588–598, APR 19 2016.

[24] L. Brenan, A. Andreev, O. Cohen, S. Pantel, A. Kamburov, D. Cacchiarelli, N. S. Persky, C. Zhu, M. Bagul, E. M. Goetz, A. B. Burgin, L. A. Garraway, G. Getz, T. S. Mikkelsen, F. Piccioni, D. E. Root, and C. M. Johannessen, "Phenotypic Characterization of a Comprehensive Set of MAPK1/ERK2 Missense Mutants," *Cell Reports*, vol. 17, pp. 1171–1183, OCT 18 2016.

[25] V. E. Gray, R. J. Hause, and D. M. Fowler, "Analysis of Large-Scale Mutagenesis Data To Assess the Impact of Single Amino Acid Substitutions," *Genetics*, vol. 207, pp. 53–61, SEP 2017.

[26] B. V. Adkar, A. Tripathi, A. Sahoo, K. Bajaj, D. Goswami, P. Chakrabarti, M. K. Swarnkar, R. S. Gokhale, and R. Varadarajan, "Protein model discrimination using mutational sensitivity derived from deep sequencing," *Structure*, vol. 20, no. 2, pp. 371–381, 2012.

[27] T. S. Wong, D. Roccatano, M. Zacharias, and U. Schwaneberg, "A statistical analysis of random mutagenesis methods used for directed protein evolution," *Journal of molecular biology*, vol. 355, no. 4, pp. 858–871, 2006.

[28] A. Currin, N. Swainston, P. J. Day, and D. B. Kell, "Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently," *Chemical Society Reviews*, vol. 44, no. 5, pp. 1172–1239, 2015.

[29] T. Abdullah, M. Faiza, P. Pant, M. R. Akhtar, and P. Pant, "An analysis of single nucleotide substitution in genetic codons-probabilities and outcomes," *Bioinformation*, vol. 12, no. 3, p. 98, 2016.

[30] S. Matuszewski, M. E. Hildebrandt, A.-H. Ghenu, J. D. Jensen, and C. Bank, "A statistical guide to the design of deep mutational scanning experiments," *Genetics*, vol. 204, no. 1, pp. 77–87, 2016.

[31] Y. Bromberg, G. Yachdav, and B. Rost, "SNAP predicts effect of mutations on protein function," *Bioinformatics*, vol. 24, pp. 2397–2398, OCT 15 2008.

[32] C. Chennubhotla and I. Bahar, “Signal propagation in proteins and relation to equilibrium fluctuations,” *PLOS Computational Biology*, vol. 3, pp. 1716–1726, SEP 2007.

[33] S. Henikoff and J. Henikoff, “Amino-acid substitution matrices from protein blocks,” *Proceedings of the National Academy of Sciences, USA*, vol. 89, pp. 10915–10919, NOV 15 1992.

[34] J. Kyte and R. F. Doolittle, “A simple method for displaying the hydropathic character of a protein,” *Journal of molecular biology*, vol. 157, no. 1, pp. 105–132, 1982.

[35] C. Sruthi and M. Prakash, “Amino acid impact factor,” *PloS one*, vol. 13, no. 6, p. e0198645, 2018.

[36] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, “Protein sectors: evolutionary units of three-dimensional structure,” *Cell*, vol. 138, no. 4, pp. 774–786, 2009.

[37] G. Csardi and T. Nepusz, “The igraph software package for complex network research,” *InterJournal, Complex Systems*, vol. 1695, no. 5, pp. 1–9, 2006.

Chapter 6

Towards Developing Intuitive Rules for Protein Variant Effect Prediction Using Deep Mutational Scan Data

Abstract

Protein structure and function can be severely altered by even a single amino acid mutation. Predictions of mutational effects using extensive artificial intelligence (AI) based models although accurate, remain as enigmatic as experimental observations in terms of improving intuitions about the contributions of various factors. Inspired by Lipinski's rules for drug-likeness, we devise simple thresholding criteria on five different descriptors such as conservation, which have so far been limited to qualitative interpretations such as high conservation implies high mutational effect. We analyse systematic deep mutational scan data of all possible single amino acid substitutions on 6 proteins to firstly define these thresholds, and then to evaluate the scope and limits of the predictions. At this stage, the approach allows us to comment easily and with a low error rate on the mutations classified as neutral or deleterious by all the descriptors, and not on the complete set of mutations. We hope that complementary to the accurate AI predictions, these thresholding rules or their subsequent modifications will serve the purpose of codifying the knowledge about the effects of mutations.

6.1 Introduction

As has been discussed in the previous chapter, deep mutational scan [1] studies have been generating unprecedented amounts of data[2–7] parallel to which computational methods have been developed to predict the mutational effect scores[8–13]. These models use tens to hundreds of variables that represent the site-specific factors, or the interactions with the immediate neighborhood. Though all these predictors may not work well for all proteins, as the detailed analysis of the predictive power of different predictors shows,[14] the performance of these AI based computational models may be considered satisfactory depending on the specific requirements, and many of these are easy to use with a web interface.[10, 11] Thus the experimental data or its computational predictions are at a stage where they can reliably generate libraries of the effects of mutations. Both these approaches are used referentially for knowing the effects of specific mutations rather than to contribute towards an understanding of the mutational landscape. However, in general, there has been a growing criticism against the lack of transparency in the AI based models that is leading to the emergence of interpretable or explainable AI.[15] The approach can be used to understand the contributions of each variable to individual predictions as is presented in Chapter 8.[16] However, even with the accurate predictions of AI, and interpretable contributions to these predictions, there is no codification of the knowledge or a reconciliation with the classical intuitions about the effects of mutations.

In the field of rational drug discovery, two very different approaches are used to screen through the leads to identify the activity or the drug-likeness. One is using highly accurate prediction models for quantitative structure activity relationships,[17] the other is using intuitive rules of thumb known as Lipinski's rules,[18] to classify the drug candidates. The latter, while not meant to be an accurate prediction of activity, is an intuitive and practically useful tool and our approach in this work is inspired by it. We revisit the qualitative intuitions on how different physico-chemical factors are independently likely to affect the function of proteins, most of which are based on site-specific descriptors such as conservation and neighborhood descriptors such as number of contacts. We ask if quantitative rules of thumb can be derived. The limitations in accuracies arising from such rules are also quantified along with these thresholds. We demonstrate the results of combining different intuitive rules for improving the reliability of predictions, albeit for a small set of mutations.

6.2 Methods

The present analyses are based on the deep mutational scan data obtained for seven proteins - β-lactamase,[19] aminoglycoside 3'-phosphotransferase (APH(3')-II),[20] heat shock protein 90 (Hsp90),[21] mitogen-activated protein kinase 1 (MAPK1),[22] ubiquitin-conjugating enzyme E2 I (UBE2I),[12] thiamin pyrophosphokinase (TPK1)[12] and β-glucosidase (Bgl3).[23] The structures of these proteins were obtained from protein data bank repository using PDB identities 1M40, 1ND4, 2CG9, 4NIF, 2UYZ, 3S4Y and

1GNX, respectively. Hydrogen atoms were added to the structure, using GROMACS.[24] Solvent accessible surface area (SASA) for each wild type residue was calculated using these structures with the *gmx sasa* tool of GROMACS[24] and a probe radius 1.4 Å. For β-lactamase and β-glucosidase homologous sequences were obtained from Pfam database[25](Pfam ID: PF13354 and PF00232 respectively) using PDB ID as the query. For other proteins, sequences obtained through PSI-BLAST search were aligned using clustal omega. The alignment was then truncated to the reference sequence and the sequences with more than 20% gaps were removed. Conservation is quantified as the frequency of highest occurring amino acid at each position in the alignment. While studying the effect of a categorical charge type change, amino acids were grouped into four categories - Positively charged (R, H, K), negatively charged (D, E), polar (S, T, N, Q, C) and hydrophobic (A, V, I, L, M, F, Y, W). P and G were not included in any group.

While analyzing the data for β-lactamase, relative-fitness,[19] $R = log_{10}\left(f^{mutant}/f^{wild-type}\right)$ where f is the ratio of allele counts in the selected and unselected population was used as a measure of phenotypic outcome of mutations. Zero, negative and positive R reflect neutral, loss of function and gain of function mutations respectively. Interestingly, two independent deep scan studies [19, 26] of β-lactamase in *E. coli* obtained a non-linear, but highly correlated outcomes (Figure E.1 in Appendix E). We chose to work with the data of Stiffler *et al.* [19] as it was 100% complete with all 19 substitutions studied for all wild type amino acids in the mutagenized region. For the proteins APH(3')-II, Hsp90 and MAPK1 the data was obtained from the study of Gray *et al.*[27] where the mutational scores are available as relative fitness (R). Fitness scores as growth rates for TPK1 and UBE2I were obtained from Weile *et al.*[12] Log$_2$ enrichment ratio for variants of Bgl3 was taken from the study of Romero et al.[23] Since the relative fitness data we used was quantitative, to perform a classification analysis, a choice of fitness threshold was required. The fitness distribution from each protein was fit to a bi-gaussian model, which is supposed to represent the neutral and deleterious mutation groups. All mutations with a fitness score more than $(\mu - 2\sigma)$ where μ, σ are the mean and standard deviation of the gaussian mode corresponding to the neutral mutations are considered neutral and others as deleterious. Unlike the case of other proteins, for MAPK1, the positive and negative scores represented deleterious and neutral mutations respectively and the choice of threshold was adapted accordingly for this data set.

6.3 Results

6.3.1 Developing thresholds for classification

We analyzed the deep mutational scan data of 7 proteins, β-lactamase, APH(3')-II, Hsp90, MAPK1, UBE2I, TPK1 and Bgl3 for which structural information as well as mutational effects data of at least 2500 substitutions are available (**Methods**). 6 of these data sets (22421 variants) were used for obtaining the thresholds and the data on Bgl3 (2732 variants) was used for an independent validation. Six physico-chemical parameters - conservation, charge type change, solvent accessible surface area (SASA), number of structural contacts, BLOSUM substitution matrix score and distance from the catalytic site (catalytic distance) were studied for identifying correlations with fitness. Each of these descriptive parameters depends either on structure, sequence or the nature of substitutions. All the parameters are site-specific, intuitive, and are widely used for inferring mutational effects. In the following sections, the fitness data of each protein was individually studied relative to each of these physico-chemical parameters. Specifically for the mutational data of β-lactamase, we studied in detail to see if correlations of the physico-chemical parameters and the deviations from them were intuitive. When it appeared that all the mutational effects could be inferred from one parameter or the other, we performed two statistical analyses of fitness relative to every parameter, Spearman correlation and F1 score. The Spearman correlation of the phenotypic outcomes relative to these individual parameters prompted us to perform thresholding relative to each of these parameters. To identify the threshold on a given parameter, we scan across the complete parameter range and use F1 score[28] to quantify the quality of classification at each value of the parameter for both the neutral and deleterious classes. $\mathrm{F1}_{\mathrm{neutral}}$ can be calculated as $\mathrm{F1}_{\mathrm{neutral}} = 2 \times (\mathrm{Precision} \times \mathrm{Recall})/(\mathrm{Precision} + \mathrm{Recall})$. Here precision is the ratio of number of true neutral predictions to total number of neutral predictions and recall is the ratio of number of true neutral predictions to number of observed neutral mutations. Similarly $\mathrm{F1}_{\mathrm{deleterious}}$ is also calculated and the threshold at which the average of F1 scores of both neutral and deleterious classes ($\mathrm{F1}_{\mathrm{avg}} = (\mathrm{F1}_{\mathrm{neutral}} + \mathrm{F1}_{\mathrm{deleterious}})/2$) is maximum was chosen as optimal. The procedure was repeated with each parameter for all proteins except Bgl3. The Bgl3 data was used as a test set for evaluating the utility of the thresholds obtained in classifying variants of a new protein.

6.3.2 Conservation threshold

Typically, evolutionary conservation reflects the functional importance of an amino acid. Figure 6.1 shows the relation between conservation and the fitness effects from deep scan data of TEM-1 β-lactamase. As suggested by the mean fitness value for a given range of conservation highlighted in Figure 6.1, there is a reduction in fitness when conserved amino acids are mutated. However, conservation alone does not clearly resolve the effect on fitness as one can see several exceptions with high fitness consequences for substitutions at poorly conserved sites and low fitness consequences at highly conserved positions. We highlight the exceptions to the expected intuitions: (1) The amino acids that have less than 20% conservation and yet severely affect function upon mutation (Relative fitness, $R <$ -1). The mutations N52C, K55(C,P), E58(C,F,H,I,L,M,P,V,W,Y), S82(C, P), S98P, N100C, T140P, T141(F,K,P,W,Y), E197(F, L), P219(F,I,W,Y), F230(C,D,E,G,I,K,L,N,P, Q,R,S,T) and S258P are deleterious even though the wild type residue is poorly conserved. All these substitutions are away from the catalytic sites and other than N52C, S82C, N100C, F230I and F230L, involved a charge type change. Interestingly, most of these substitutions also involved a loss of solubility [29] which could be the reason for reduced functional fitness. (2) The amino acids were conserved ($>$ 80%) but their substitution did not affect the function significantly. G156D, G156E, G156N and G236A are the substitutions which are neutral despite high conservation. Also in these cases the wild type amino acid is substituted with amino acid of different charge type. As conservation quantifies only variability at a specific position and does not distinguish different substitutions, we calculated position specific scoring matrix (PSSM) using PSI-BLAST and explored its relation with fitness. We observe only a weak correlation (Figure E.2).

We further attempted to quantify a threshold for the general intuition that higher the conservation, greater are the fitness consequences of substituting it. We scanned across for different values of the threshold and quantified the F1 score for both neutral and deleterious classes (Figure 6.1). The same analysis performed for the other five proteins is shown in Figure E.3 of Appendix E. It can be seen that the intuition holds for all proteins as indicated by the change in the mean fitness with conservation. The optimal threshold for conservation according to the F1 score was 0.35 for β-lactamase. For the other five proteins we studied the threshold varied in the range 0.45 to 0.9 (Table 6.1).

6.3.3 Solvent accessible surface area (SASA) threshold

The relation between fitness and SASA of wild type amino acid which reflects how buried the amino acids are is shown in Figure 6.2 (alanine scan results in Figure E.4 of Appendix E). The intuitive learning from this figure is that substitutions at amino acids which are completely buried can potentially range from neutral to deleterious, while the effect tapers off for amino acids with high SASA values which have minimal effect on fitness. The mutations defining the frontier and showing the highest fitness compromise at any given SASA, were recorded by taking note of the alanine scan mutations near the triangular border in the plot. Of the amino acids P27, L57, R61, R65, F66, S70, K73,

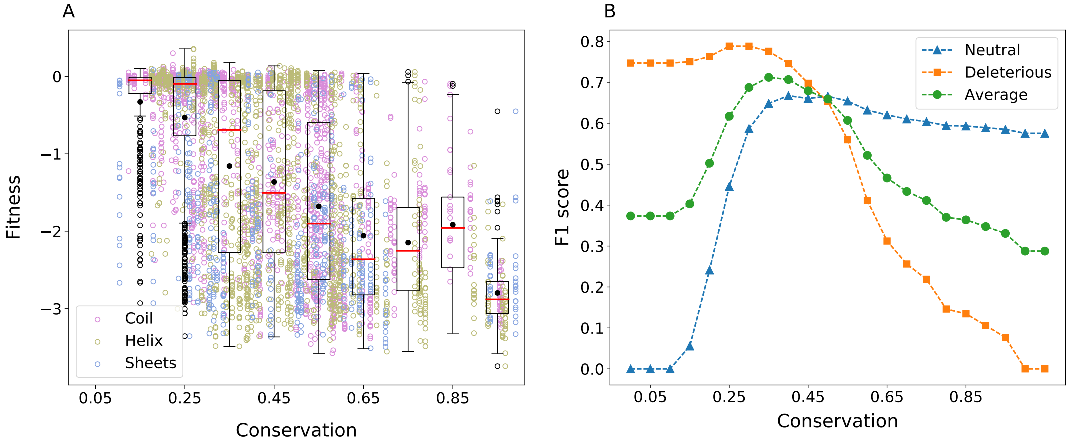

Figure 6.1: Effect of conservation. (A) The relationship between conservation and fitness was studied using the homologous sequences for TEM-1 β-lactamase (Pfam ID PF13354). It can be seen that the number of neutral substitutions decreases considerably for amino acids with conservation > 60%. The black filled circle and the red line represent mean and median of the fitness respectively. The whiskers are plotted at 1.5 times the interquartile range and black open circles show the outliers. (B) Changes in F1 score for the neutral and deleterious classes and the average of both plotted as the threshold for conservation to classify the mutations is varied.

R93, Y105, S130, N132, N136, D157, R161, E166, R222, W229 and W290, which are on the frontier of highest fitness loss, most are near the binding pocket. W229 is known to have an allosteric effect on the function.[30] However, the reasons for the functional compromise of mutations at P27 and R222 are not clear. The data on average supports the intuition that amino acids which are completely buried and have a zero or reduced solvent accessible area do not tolerate mutations. At intermediate solvent accessibility conditions, interestingly, a reduction in volume of the amino acid seems to be more deleterious in general. This could be because of the cavities being created which affects the packing of the residues. It is known that cavity creating mutations reduce the stability of proteins.[31] Applying a thresholding condition on SASA that classifies the fitness consequences of mutations as neutral or deleterious, we obtain 0.3 nm^2 as the optimal threshold (Figure 6.2 and Table 6.1). For other proteins the optimal threshold for SASA was observed to be in the range 0.1 to 0.4 nm^2 (Table 6.1). The fitness distributions at different ranges of SASA for these proteins are given in Figure E.5 of Appendix E. The distributions for APH(3')-II, Hsp90 and MAPK1 follow similar trend as seen in the case of β-lactamase and for TPK1 and UBE2I there is comparatively higher variability in fitness even at lower solvent accessibility.

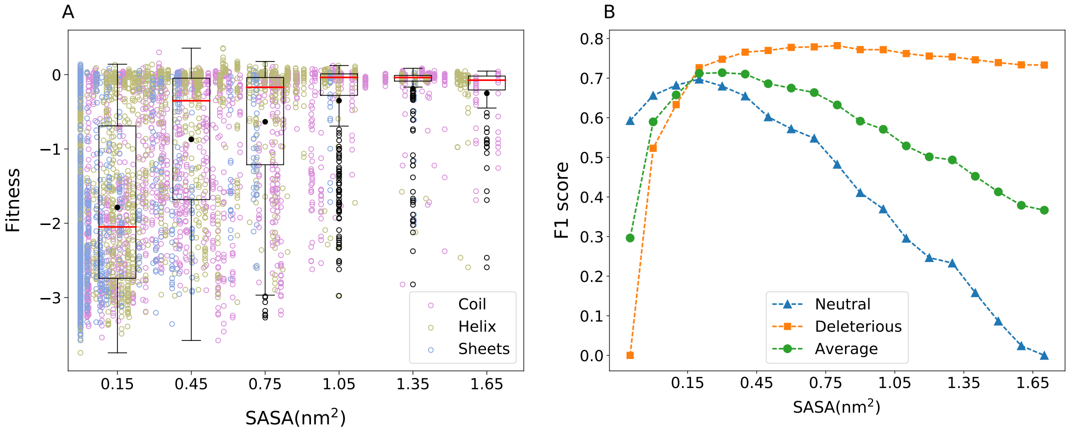

Figure 6.2: Effect of solvent accessibility. (A) Solvent accessibility for all amino acids of β-lactamase was calculated using the 3D protein structure (PDB ID: 1M40). SASA versus fitness shows a half-triangular pattern. The deviations from this half-triangular pattern are noted in the main text. For details about the box plot representation, see Figure 6.1. (B) Substitutions are classified as neutral and deleterious based on a chosen SASA threshold and the quality of resulting classification is quantified using F1 score. F1 scores when different SASA thresholds are used is shown.

6.3.4 Threshold for number of inter-residue contacts

Inter-amino acid interactions mediated by hydrogen bonds, salt bridges, stackings etc determine how much a substitution disturbs the overall structural stability and function. While the biochemical details of the different interactions may be explored, and whether or not the nature of the substitutions conform with the existing interactions may also be investigated, a simpler metric is the total number of inter-residue interactions any given residue is involved in. We studied this by counting the number of atom level interactions that an amino acid is involved in, and the sensitivity to its substitution. We used the native structure of the protein obtained from the protein data bank and along with an interaction cutoff of 4 Å to count interactions. Figure E.6 of Appendix E shows that the relation between fitness and number contacts is weak. While the average trends are intuitive, like substitutions of residues with higher number of contacts result in larger fitness effect, the variation in fitness for a given number of contacts is high. An optimal threshold for the number of inter-residue contacts was found to be 14 for β-lactamase. The F1 score variation with respect to number of contacts for other proteins and the optimum thresholds obtained are given in Figure E.7 of Appendix E and Table

6.1 respectively. There is an intuitive monotonous variation in mean fitness with number of contacts for the cases of APH(3')-II and UBE2I whereas for Hsp90, TPK1 and MAPK1 the fitness changes do not seem to have dependence on the number of contacts.

6.3.5 BLOSUM threshold

All other physico-chemical metrics mentioned so far depend on the wild type amino acid alone, and do not reflect the nature of the substitution. We use BLOSUM65 matrix which statistically summarizes the naturally occurring substitution probabilities across all proteins to see if the fitness effects of an amino acid substitution can be captured by it. A plot of BLOSUM score of substitutions and their fitness effects in β-lactamase are shown in Figure 6.3. One can also infer an optimal threshold for the BLOSUM for β-lactamase from this. The dependence of fitness on BLOSUM matrix score can be seen for all proteins in Figure E.8 of Appendix E.

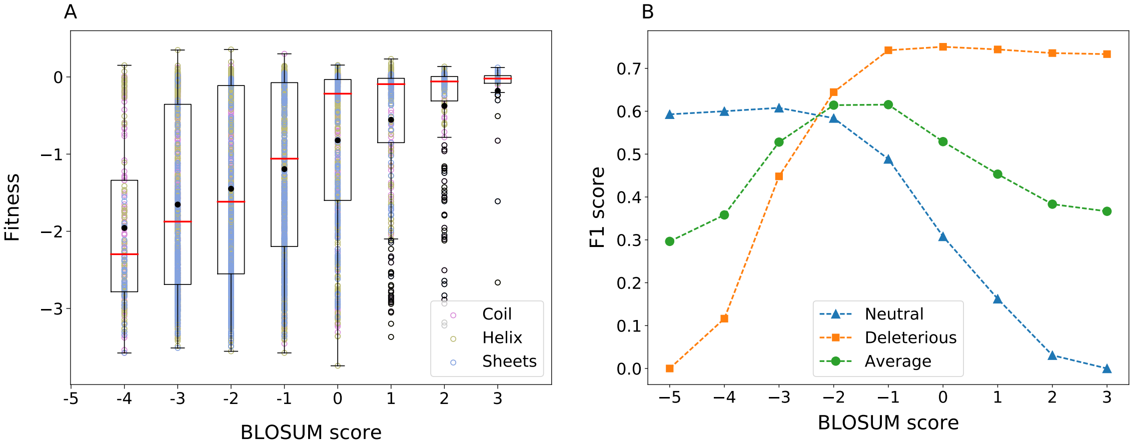

Figure 6.3: Effect of BLOSUM substitution score. (A) Fitness scores for substitutions in β-lactamase as a function of the BLOSUM62 score for the substitution. See Figure 6.1 for details about the box plot representation. (B) F1 score when each of the substitution matrix score is used as a threshold to classify mutations as neutral and deleterious. Average of F1 scores of both classes is also plotted.

6.3.6 Charge-invariant fitness map

Another physical intuition about the nature of the substitutions is that charge type changes can disrupt local interactions or solvent accessibility and lead to a loss of structure and function. Four amino acid categories were considered - positively charged, negatively

Protein name *Parameter*	β-lactamase	APH(3')-II	Hsp90	MAPK1	UBE2I	TPK1	Average threshold
Fitness cut-off	-0.5	-2.5	-0.3	0.5	0.2	0.4	-
Conservation	0.35	0.5	0.85	0.9	0.45	0.55	0.6
SASA (nm^2)	0.3	0.2	0.2	0.1	0.2	0.4	0.2
Contacts	14	19	25	20	18	14	18
BLOSUM	-1	-3	-3	-3	-2	-1	-2

Table 6.1: Threshold for parameters. Thresholds for different parameters were obtained by maximizing F1$_{\text{avg}}$ which is the average of F1 scores of neutral and deleterious class predictions. The thresholds were obtained for each parameter and data from every single protein. For a parameter, the average of thresholds obtained for the 6 proteins was calculated to obtain the average threshold.

charged, polar and hydrophobic (**Methods section**). In an attempt to highlight the functional effects that are not intuitively expected, we present the analysis only for the mutations where the charge type of the mutant is the same as the charge type of the wild type, yet the mutation causes a severe loss in fitness (Figure 6.4). Since charge type change is categorical in nature, the consequences of this prediction can be summarized in a contingency table, rather than a parameterized dependence as: 2401 true positives, 569 true negatives, 1536 false positives and 491 false negatives.

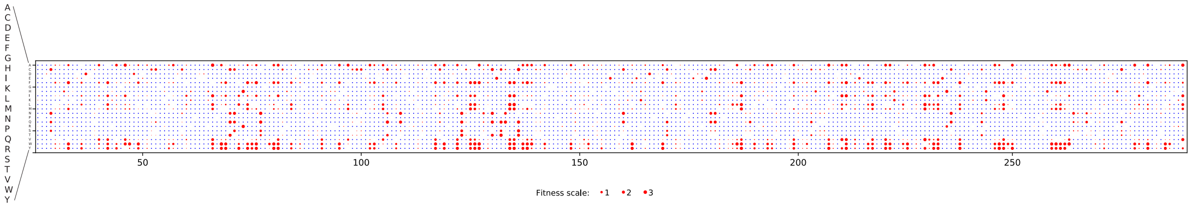

Figure 6.4: Effect of charge variation. The mutational effect scores are shown in a two-dimensional matrix representation with each row representing the amino acid substituted by and each column the position along the amino acid sequence of β-lactamase. All substitutions with no change in the charge type of the amino acid are highlighted in red filled circles with the size of the circle representing the fitness score and others in open blue circles of equal size. There are many substitutions for which the fitness is heavily compromised even with no change in the charge type.

6.3.7 Threshold for distance from catalytic site

Substitution of catalytic and binding pocket residues: Substitution of catalytic residues are expected to be mostly deleterious, which is also the reason the conservation is usually correlated to the distance from the catalytic site.[32] Any substitution in the five reported catalytic residues in β-lactamase - S70, K73, S130, E166 and A237, other than A237S and A237G leads to high fitness compromise. In catalysis, the backbones of S70 and A237 in conjunction form an oxyanion hole stabilizing a reaction intermediate, [33] thus tolerating some side chain substitutions at A237. The intolerance of all substitutions except of S and G could probably be because of size constraints. In addition to the catalytic sites noted above, residues M69, Y105, N132, N170, K234, S235, G236, G238, E240, M272, form the binding pocket. Among these, N132 and K234 are the most sensitive ones as all 19 mutations at these positions result in reduced fitness ($R <$ -0.5). **Substitution of distal amino acids:** From all mutations that lead to a loss of function [19], the mutations which also were independently seen to lead to a loss of solubility [29] were eliminated. 57 substitutions of 16 wild type amino acids were more than 1.5 nm, away from the catalytic residues and yet lead to $R < -1.5$. All these substitutions are either buried (SASA < 0.3 nm^2) or have higher inter-residue contacts (> 15) except for two which are evolutionarily not favoured (BLOSUM < 0). It is also possible that the substitutions had long range effects such has been observed in the case of some other proteins.[34, 35] The fitness effects of all the amino acid mutations studied in β-lactamase are summarized as a function of the distance from the catalytic site in Figure 6.5. 1.5 nm as a threshold distance from the catalytic residues to classify the effects of mutations optimized the true and false positive predictions.

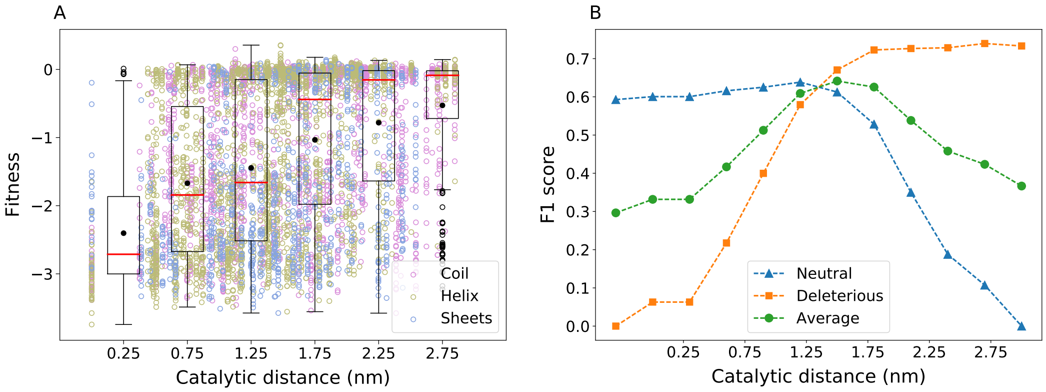

Figure 6.5: Effect of catalytic distance. (A) Fitness changes are shown with respect to the distance between wild-type residue and the closest catalytic residue. See Figure 6.1 for details about the box plot representation. (B) F1 score for the neutral, deleterious classes and the average of both are plotted as the catalytic distance threshold for classification is varied.

6.4 Discussion

6.4.1 From intuitions to thresholds

Conservation of an amino acid has been a traditional benchmark to understand the functional relevance of amino acids as well as to infer the potential effects of their mutations. The intuitions such as when the conservation of an amino acid is sufficiently high, the chance of its mutation affecting the function is also high were developed either from mutational studies or by comparing homologous proteins with sequence alignments. In this intuitive classification, two aspects remain qualitative, how high the conservation should be for it to be important and the quality of the resulting classification. Technically this information may be derived by compiling the data on all the mutations available, but has not been done to our knowledge. However, the variations across proteins, experiments make comparisons difficult. The present work uses the publicly available systematic large data sets on mutational effects to shed light on both these aspects for six proteins. The thresholds obtained for the six proteins are all summarized in Table 6.1. It appears that the conservation threshold optimizing the false positives and false negatives vary widely from 0.35 to 0.9 for different proteins. The question then arises whether the thresholds vary with larger data sets, or if one can identify universal thresholds. To address this question, one has to work with relatively large data sets, with reliable quantitative measurements of the mutational effects. At this stage, the deep mutational scan measurements with different assay conditions and varying levels of stressor concentrations are indicative of the overall trends rather than precise measurements. Since to the best of our knowledge the thresholds were not defined so far, and we suggest the use of averages of the thresholds obtained from the different proteins and defer the universality aspect until a later occasion.

6.4.2 Optimization has to balance several factors

Any classification method has to balance between true and false positives, and is likely to be biased by the over representation of the neutral or deleterious in the training. The same is true for the rules of thumb we developed. We chose $\mathrm{F1}_{\mathrm{avg}}$ score as the measure to quantify this balance. But as F1 score focuses only on one class we decided the optimal threshold as the one which maximizes the average of $\mathrm{F1}_{\mathrm{neutral}}$ and $\mathrm{F1}_{\mathrm{deleterious}}$. It is clear from the data that there is no clear parameter that can be used as a threshold or a rule of thumb for improving the true positives, without also increasing the false-positives. The false-positives and false-negatives were both minimized simultaneously by using the sum of $\mathrm{F1}_{\mathrm{neutral}}$ and $\mathrm{F1}_{\mathrm{deleterious}}$ scores. Because of this optimization, any threshold or rule based classification will be partly incorrect. Further, the data we used from the deep mutational scan has a larger fraction of neutral mutations (64%). Thus, there may be unavoidable biases in making predictions, which may be addressed with larger data in the future. Not withstanding this limitation, we probed in detail the exceptions to the thresholding rules for β-lactamase. What appears to be an exception to the monotonous relation between conservation and fitness was explained by a charge type change. The ambiguity in the sensitivity of amino acids which are in the intermediate ranges of SASA could be clarified by the change in volume upon mutation. Thus, while each of the physico-chemical parameters is not complete, they may complement each other. This is expected, since proteins are complicated, and predicting structural or functional consequences of mutations with a single biophysical or biochemical parameter is non-trivial.

6.4.3 Multifactorial classification

Before attempting a multifactorial classification based on thresholding, we developed a logistic regression model by training on 70% of the data from the six proteins. The probability of a substitution being deleterious (P_{del}) we obtained was:

$\mathrm{P_{del}} = 1/[1 + \exp(1.99 - 1.97 \times \mathrm{Conservation} - 0.014 \times \mathrm{Contacts} + 0.72 \times \mathrm{SASA} + 0.33 \times \mathrm{BLOSUM} + 0.12 \times \mathrm{Q_c})]$

where Q_c represents a charge type change, 1 if there is no change and 0 otherwise. The model was tested on the remaining 30% and had an accuracy of 0.70, better than the accuracy from the individual variables. Taking cue from this, a threshold based classification with several biochemical intuitions was systematically explored in Figure 6.6. In the analysis, a subset of mutations (2892 of them) from β-lactamase which result in a fitness compromise ($R <= -0.5$) were analyzed. Each mutation was independently classified as neutral or deleterious using different descriptors and their corresponding threshold values. The numbers appearing in the different overlap regions in Figure 6.6 indicate the number of mutations for which the variables defining the overlap all result in a false-neutral prediction. The interesting region in the center shows that when all 5 variables classify a mutation as neutral there are only 2 false predictions. In addition to this combined representation, one can also see how the number of false-neutral predictions reduces as the number of descriptors are increased one after another (Figures E.9 and E.10 of Appendix E). Of course, in this approach now a new kind of uncertainty will remain for a fraction of the substitutions when about half of the descriptors suggest a deleterious effect and others point to neutrality.

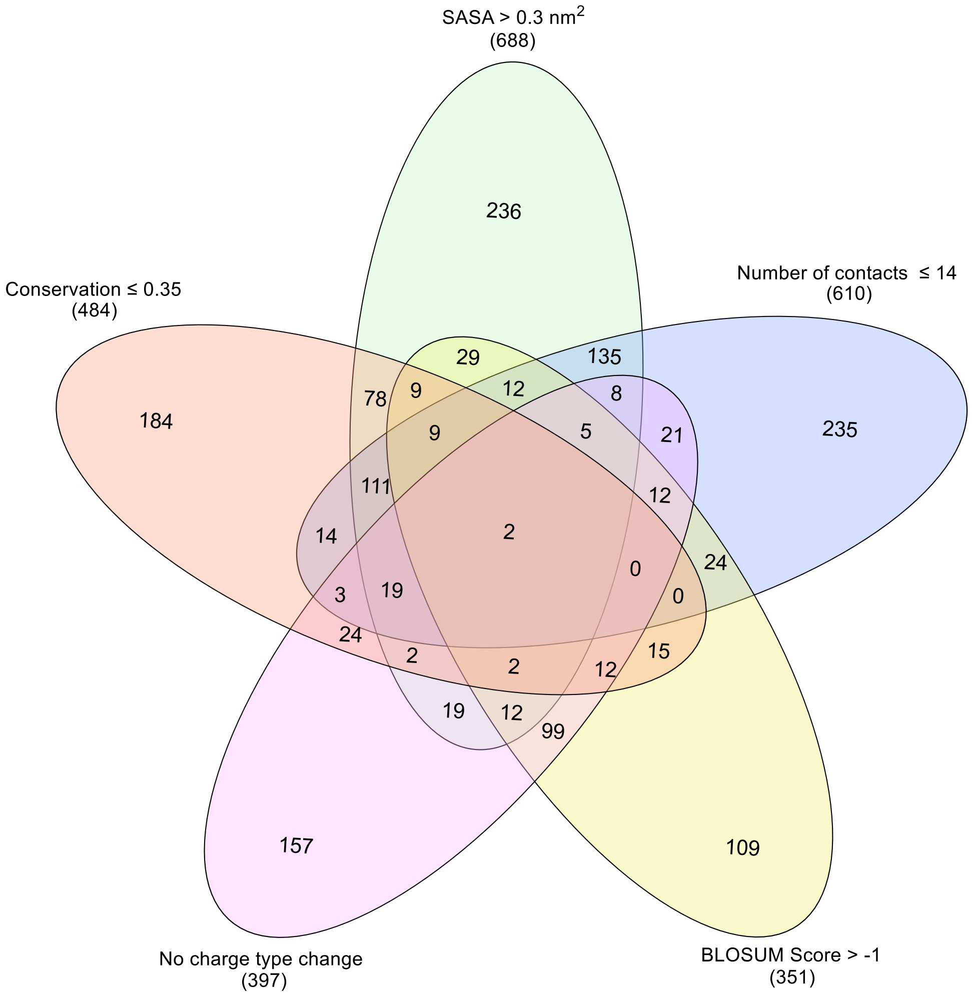

Figure 6.6: Reducing false predictions by combining parameters. Venn diagram showing the number of substitutions that do not follow the intuitions related to different structural and sequence related properties of the wild type and the substituting amino acids. Thresholds indicated in the figure were used on each of the individual parameters, to classify the mutations as deleterious or neutral. The central region indicates that there are only 2 false-neutral predictions when all five variables classify the mutation as neutral. The number of total false neutral predictions when only one variable is used is given in brackets below the variable labels.

6.4.4 Common thresholds for many proteins

It is clear that the thresholds of the physico-chemical parameters that we obtained for different proteins varied significantly. For each protein, when multiple threshold criteria were satisfied the error rate in those cases was smaller. We asked if it is possible to define universal thresholds or at least common thresholds for the data sets we have studied. For each of the physico-chemical parameters, we used an average derived from the different proteins, i.e., the row averages in the Table 6.1. By using these averages as thresholds, we re-calculated how the error rate drops for all six proteins as shown in Figure E.11 of Appendix E. The results are encouraging, at this stage and suggest that by qualifying for at least three conditions with the threshold criterion, the chance of false-predictions drops significantly. The details of true and false, neutral and deleterious classifications when these average thresholds are applied individually on each protein are given in Table E.1 of Appendix E. The classification based on average thresholds performs better for variants which are suggested as either neutral or deleterious by all variables compared to the classification based on BLOSUM substitution score as variants with score < 0 considered as deleterious and all others as neutral (0.76 versus 0.66, McNemar's p-value $=0$). The average thresholds were tested on an independent data set of β-glucosidase (Bgl3)[23] that was not used for obtaining the common thresholds. Interestingly, the fraction of false neutral and false deleterious obtained are low, 0.07 and 0.16 respectively. We also performed F1 score analysis by combining data sets of all proteins and defined a common threshold across proteins for each parameter. The thresholds obtained from this analysis were same as the average threshold for number of contacts, SASA and BLOSUM and for conservation a threshold of 0.5 was obtained. As there is only change in the conservation threshold, we did not repeat any other analyses with these set of thresholds. Leave one protein out analysis in which the thresholds obtained for five proteins are averaged and the quality of predictions are tested on the data set of the sixth protein was performed. In all cases the quality of predictions was similar to that as with the average of thresholds for all six proteins (Table E.2). We also explored the variation in the thresholds on using another criterion for threshold determination such as maximizing the difference between True Positive Rate (TPR) and False Positive Rate (FPR) as obtained from an ROC analysis. While there were small changes in the threshold for individual proteins, the common threshold for number of contacts and BLOSUM score remained the same. For conservation and SASA the thresholds changed from 0.6 to 0.5 and 0.2 to 0.3 respectively. The effect of small variations such as this is discussed in the Section 6.4.6.

6.4.5 Effect of experimental error on the thresholds

We wanted to check whether the thresholds identified are robust against the uncertainty in the experimentally determined fitness scores, using the measurement errors available for β-lactamase, TPK1 and UBE2I. In the present work, any variant is first classified as neutral or deleterious by checking its fitness relative to a bi-gaussian distribution, and then the predictive capacity of the variable was evaluated. However, for some of the variants the two extremes for their fitness accounting for the error (variant fitness - standard deviation and variant fitness + standard deviation) can suggest different classifications, creating an uncertainty in what was meant to be a reference. We eliminated these uncertain variants from our analysis and recalculated the thresholds for the variables. For TPK1 eliminating these variants did not change any threshold, for β-lactamase the BLOSUM threshold changed from -1 to -2 and for UBE2I the threshold on the number of contacts changed from 18 to 22. However, while averaging to obtain the common thresholds, only the average threshold for number of contacts changed from 18 to 19. This variation was within the scope of the sensitivity analysis we perform in Section 6.4.6 below, and practically the threshold may be considered robust relative to the experimental errors.

6.4.6 Sensitivity analysis

Since the thresholds from different proteins varied, and the data was not sufficient to comment on universal thresholds, we performed a sensitivity analysis for the qualitative changes in conclusions with small changes in the thresholds. We varied the average thresholds (th_{av}) by an amount δth which is approximately equal to 10% of the maximum value for that parameter. We quantified the fraction of wrong predictions for both the neutral and deleterious classes when each parameter was varied, one at a time, as $(th_{av} \pm \delta th)$. We find that for the mutations which are predicted as neutral or deleterious by all five variables, this fraction remains same in most of the cases though there are differences in the number of mutations identified on changing the threshold. (Table E.3 of Appendix E)

6.4.7 Scope of present work in the context of existing AI predictors

Advances in artificial intelligence are leading discoveries in several areas of science and engineering. The same is true for protein effects predictions, where models such as SNAP[10], Envision[11] or others[12] continue to improve the accuracy of classifications or fitness predictions. Many of the models have even created a easy to use web based interface. The SNAP predictions are analyzed from this perspective by choosing only those mutations which were classified with an expected accuracy greater than 80%. The number of neutral or deleterious predictions we obtain with the average thresholds listed in Table 6.1 are also shown in Table 6.2. It can be seen that though for a smaller set, the fraction of false predictions when thresholding criteria are used is comparable to that of SNAP. Clearly, when considering the complete set of mutations, the present work is no match to the AI models. However, in several areas of AI, there has been a concern about the lack of transparency in the way AI treats the predictions, with a two-fold motivation: 1. assuming the predictions are correct, is it possible to find the contributions to each individual prediction so that one has a better understanding of the final output. For example, such factor contributions can help find mutations where the solubility and fitness changes from the mutation are both acceptable. 2. Although on average the prediction quality is high, how can one be sure that a specific prediction is reliable? Is it possible to, for example, correlate the factor contributions, with known intuitions so that one gains confidence in the final prediction of the effects? Thus, there is always a need for intuitions or at least rules of thumb which empirically codify the observations. Further, despite the ability to have accurate calculations, in other parts of the literature the qualitative statements about critical parameters being high or low continue to exist. The present work, while acknowledging its shortcomings relative to the AI based models aims to improve the qualitative intuitions by quantifying them with thresholds. Of course, the limitations are that it is not easy to comment on mutations where the suggestions from the five parameters are not correlated, and it is possible to comment only on mutations where all the parameters suggest the mutation to be neutral or deleterious. Given this limitation, when a mutation is implicated for a disease, for example, the present method is not useful for classifying such critical mutations. Instead it can be used as an inverse approach where the mutations suggested by the thresholds to be deleterious or neutral can be believed to be so with a high degree of accuracy. The limitation of the thresholding could be related to the small training set or the site-specific nature of the descriptors we chose which may fail to capture distal effects of mutations. The present approach is an important step toward codifying the learnings from a pedagogical perspective and is helpful for quick analysis when all parameters suggest a similar outcome. The question of whether these thresholds can be universal is beyond the scope of the present work and should be revisited with data much larger than what has been used and descriptors that capture long range effects of mutations.

In line with the idea of having a simple, interpretable model for mutational effect prediction, as presented in this Chapter, in the coming chapters we attempt to develop models that have a transparent decision process but are more accurate than the thresholds.

6.5 Conclusions

Different physico-chemical factors describing native amino acids or their substitutions were evaluated in this work for their potential to capture the loss of protein function upon mutation. Visual representations of the large scale mutational data on six proteins were used to establish correlations, albeit weak ones, between the individual descriptors and the functional effects. We attempted to obtain a double quantification of the common intuitions such as when the descriptor is sufficiently large it is likely to have a significant effect. Threshold values for the descriptors which can be used for classification and the

	SNAP predictions with expected accuracy $\geq$ 80%				Using averaged thresholds			
Protein	Number of mutations predicted as neutral	Fraction of false neutral predictions	Number of mutations predicted as deleterious	Fraction of false deleterious predictions	Number of mutations predicted as neutral	Fraction of false neutral predictions	Number of mutations predicted as deleterious	Fraction of false deleterious predictions
β-lactamase	802	0.13	1317	0.04	200	0.18	183	0
APH(3')-II	1156	0.03	673	0.47	289	0.04	186	0.45
Hsp90	578	0.01	1151	0.59	231	0.02	260	0.58
MAPK1	594	0.02	1507	0.44	185	0.03	447	0.40
UBE2I	32	0.06	751	0.26	146	0.14	161	0.11
TPK1	655	0.41	668	0.31	183	0.37	207	0.31
Bgl3§	665	0.14	695	0.17	140	0.07	190	0.16

Table 6.2: SNAP predictions with expected accuracy $\geq$ 80% were selected and the fraction of false neutral/deleterious predictions for this set of mutations were calculated. Using the average of thresholds tabulated in Table 1, mutations predicted as neutral or deleterious by all five variables were identified and the fraction of false predictions for these sets are given in the table. It can be seen that the quality of classification achieved using just simple thresholding criteria compare with that of SNAP though for a smaller set of mutations. §Bgl3 data which was not used for training, was added as an independent validation

consequent false predictions were discussed. Combination of these simple rules of thumb improves the confidence in the predictions, although of a smaller set of mutations. The approach thus attempts to quantify the physico-chemical intuitions, which we believe is complementary to the more accurate but complex machine learning based approaches.

Bibliography

[1] D. M. Fowler and S. Fields, "Deep mutational scanning: a new style of protein science," *Nature Methods*, vol. 11, no. 8, p. 801, 2014.

[2] V. E. Gray, K. Sitko, F. Z. N. Kameni, M. Williamson, J. J. Stephany, N. Hasle, and D. M. Fowler, "Elucidating the molecular determinants of aβ aggregation with deep mutational scanning," *bioRxiv*, p. 662213, 2019.

[3] J. M. Lee, J. Huddleston, M. B. Doud, K. A. Hooper, N. C. Wu, T. Bedford, and J. D. Bloom, "Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human h3n2 influenza variants," *Proceedings of the National Academy of Sciences*, vol. 115, no. 35, pp. E8276–E8285, 2018.

[4] E. Jones, N. Lubock, D. Cancilla, M. Satyadi, R. Jajoo, and S. Kosuri, "Deep mutational scanning of the beta-2 adrenergic receptor," in *PROTEIN SCIENCE*, vol. 27, pp. 58–58, WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 2018.

[5] J. Weile and F. P. Roth, "Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas," *Human genetics*, vol. 137, no. 9, pp. 665–678, 2018.

[6] J. B. Kinney and D. M. McCandlish, "Massively parallel assays and quantitative sequence–function relationships," *Annual review of genomics and human genetics*, 2019.

[7] A. Nisthal, C. Y. Wang, M. L. Ary, and S. L. Mayo, "Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis," *Proceedings of the National Academy of Sciences*, vol. 116, no. 33, pp. 16367–16377, 2019.

[8] T. A. Hopf, J. B. Ingraham, F. J. Poelwijk, C. P. I. Scharfe, M. Springer, C. Sander, and D. S. Marks, "Mutation effects predicted from sequence co-variation," *Nature Biotechnology*, vol. 35, pp. 128–135, FEB 2017.

[9] A. J. Riesselman, J. B. Ingraham, and D. S. Marks, "Deep generative models of genetic variation capture the effects of mutations," *Nature Methods*, vol. 15, pp. 816+, OCT 2018.

[10] Y. Bromberg, G. Yachdav, and B. Rost, "SNAP predicts effect of mutations on protein function," *Bioinformatics*, vol. 24, pp. 2397–2398, OCT 15 2008.

[11] V. E. Gray, R. J. Hause, J. Luebeck, J. Shendure, and D. M. Fowler, "Quantitative missense variant effect prediction using large-scale mutagenesis data," *Cell Systems*, vol. 6, no. 1, pp. 116–124, 2018.

[12] J. Weile, S. Sun, A. G. Cote, J. Knapp, M. Verby, J. C. Mellor, Y. Wu, C. Pons, C. Wong, N. van Lieshout, F. Yang, M. Tasan, G. Tan, S. Yang, D. M. Fowler, R. Nussbaum, J. D. Bloom, M. Vidal, D. E. Hill, P. Aloy, and F. P. Roth, "A framework for exhaustively mapping functional missense variants," *Molecular Systems Biology*, vol. 13, DEC 2017.

[13] C. K. Sruthi and M. K. Prakash, "Deep2full: Predictive model for complementing phenotypic outcomes in a deep mutational scan using protein sequence and structure information," *bioRxiv*, 2017.

[14] B. J. Livesey and J. A. Marsh, "Using deep mutational scanning data to benchmark computational phenotype predictors and identify pathogenic missense mutations," *BioRxiv*, p. 855957, 2019.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

[16] C. Sruthi and M. K. Prakash, "Interpreting mutational effects predictions, one substitution at a time," *bioRxiv*, p. 867812, 2019.

[17] R. Kadam and N. Roy, "Recent trends in drug-likeness prediction: a comprehensive review of in silico methods," *Indian Journal of Pharmaceutical Sciences*, vol. 69, no. 5, p. 609, 2007.

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

[19] M. A. Stiffler, D. R. Hekstra, and R. Ranganathan, "Evolvability as a function of purifying selection in tem-1 β-lactamase," *Cell*, vol. 160, no. 5, pp. 882–892, 2015.

[20] A. Melnikov, P. Rogov, L. Wang, A. Gnirke, and T. S. Mikkelsen, "Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes," *Nucleic Acids Research*, vol. 42, no. 14, 2014.

[21] P. Mishra, J. M. Flynn, T. N. Starr, and D. N. A. Bolon, “Systematic Mutant Analyses Elucidate General and Client-Specific Aspects of Hsp90 Function,” *Cell Reports*, vol. 15, pp. 588–598, APR 19 2016.

[22] L. Brenan, A. Andreev, O. Cohen, S. Pantel, A. Kamburov, D. Cacchiarelli, N. S. Persky, C. Zhu, M. Bagul, E. M. Goetz, A. B. Burgin, L. A. Garraway, G. Getz, T. S. Mikkelsen, F. Piccioni, D. E. Root, and C. M. Johannessen, “Phenotypic Characterization of a Comprehensive Set of MAPK1/ERK2 Missense Mutants,” *Cell Reports*, vol. 17, pp. 1171–1183, OCT 18 2016.

[23] P. A. Romero, T. M. Tran, and A. R. Abate, “Dissecting enzyme function with microfluidic-based deep mutational scanning,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 23, pp. 7159–7164, 2015.

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

[25] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, E. L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S. C. Tosatto, and R. D. Finn, “The Pfam protein families database in 2019,” *Nucleic Acids Research*, vol. 47, pp. D427–D432, 10 2018.

[26] E. Firnberg, J. W. Labonte, J. J. Gray, and M. Ostermeier, “A comprehensive, high-resolution map of a gene’s fitness landscape,” *Molecular Biology and Evolution*, vol. 31, no. 6, pp. 1581–1592, 2014.

[27] V. E. Gray, R. J. Hause, and D. M. Fowler, “Analysis of large-scale mutagenesis data to assess the impact of single amino acid substitutions,” *Genetics*, pp. genetics–300064, 2017.

[28] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*, vol. 463. ACM press New York, 1999.

[29] J. R. Klesmith, J.-P. Bacik, E. E. Wrenbeck, R. Michalczyk, and T. A. Whitehead, “Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning,” *Proceedings of the National Academy of Sciences*, vol. 114, no. 9, pp. 2265–2270, 2017.

[30] F. G. Avci, F. E. Altinisik, D. Vardar Ulu, E. Ozkirimli Olmez, and B. Sariyar Akbulut, “An evolutionarily conserved allosteric site modulates beta-lactamase activity,”

Journal of enzyme inhibition and medicinal chemistry, vol. 31, no. sup3, pp. 33–40, 2016.

[31] A. E. Eriksson, W. A. Baase, X.-J. Zhang, D. W. Heinz, M. Blaber, E. P. Baldwin, and B. W. Matthews, "Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect," *Science*, vol. 255, no. 5041, pp. 178–183, 1992.

[32] B. R. Jack, A. G. Meyer, J. Echave, and C. O. Wilke, "Functional sites induce long-range evolutionary constraints in enzymes," *PLoS biology*, vol. 14, no. 5, p. e1002452, 2016.

[33] B. P. Atanasov, D. Mustafi, and M. W. Makinen, "Protonation of the β-lactam nitrogen is the trigger event in the catalytic action of class a β-lactamases," *Proceedings of the National Academy of Sciences*, vol. 97, no. 7, pp. 3160–3165, 2000.

[34] N. Rajasekaran, S. Suresh, S. Gopi, K. Raman, and A. N. Naganathan, "A general mechanism for the propagation of mutational effects in proteins," *Biochemistry*, vol. 56, no. 1, pp. 294–305, 2016.

[35] N. Rajasekaran, A. Sekhar, and A. N. Naganathan, "A universal pattern in the percolation and dissipation of protein structural perturbations," *The journal of physical chemistry letters*, vol. 8, no. 19, pp. 4779–4784, 2017.

Chapter 7

A4 Size Decision Tree

Abstract

Understanding how amino acid substitutions can affect protein function is important in basic biology as well as in protein engineering. With the availability of more data and advanced artificial intelligence (AI) models, it is becoming possible to make reliable predictions of the mutational effects. However, to our knowledge the philosophical shift from predictability to interpretability has not caught on yet in mutational effects predictions. In this work, we build a decision tree (AI model) and condense it into a compact representation. The method, while approximate, is intended to transfer the power to make decisions from computers to humans, thereby allowing the possibility of learning about the factors that contribute to the mutational effects, with lesser abstraction than black box model.

7.1 Introduction

The rules of thumb developed for mutational effect prediction in the previous chapter are based on intuitions about the correlation between different properties of the wild type amino acid or the substitution and mutational effect. While the extreme cases of these variables may be easy to interpret, the classification is otherwise fuzzy with no thresholds or clear rules for separating the neutral from the deleterious effects. The other completely different approach for prediction is by using computational models that use artificial neural networks, random forests or decision trees[1] such as Deep2Full presented in Chapter 5. But as has been mentioned before, these models are trained on tens to hundreds of thousands of mutations, and use tens to hundreds of predictive variables. Some of the models such as SNAP[2] give predictive scores as well as expected accuracy and combining these the two gives good predictions of the effects. However, one shortcoming of depending on these approaches, which had been true in general for artificial intelligence (AI) based models excepting the newer ones on interpretable AI, is that while the end result may be predicted, one does not develop any insights into the predictions. For example, decision trees are common in several areas of data analysis. The implicit assumption is that the decisions are taken by computers, and the emphasis is typically on constructing the most detailed tree which may be humanly unreadable, or construct ensembles of such trees which make it even more difficult to track the process. To the best of our knowledge, either in other areas of data sciences or in mutational effects predictions, there has been no effort to simplify the decision trees to achieve a transparency in the decision process. Here, we attempt to find a middle ground by developing an AI based decision tree model, but presenting it in a compact format with a goal to make the decision process of classification transparent.

7.2 Methods

To develop a simple decision tree for neutral/deleterious classification, we used the mutational effects data from systematic and extensive deep mutational scan studies. Around 22,000 mutational effect scores (14445 neutral and 7976 deleterious mutations) from six proteins (β-lactamase,[6] APH(3')-II,[7] Hsp90,[8] MAPK1,[9] UBE2I,[10] and TPK1[10]) were collated. The classification of variants to neutral and deleterious was done as described in the Methods section of Chapter 6. Five different physico-chemical parameters characterizing the wild type amino acids and their substitutions were calculated - conservation, solvent accessible surface area (SASA), number of atom-atom contacts, BLOSUM65 substitution matrix score,[11] and by noting whether a mutation involves a change in charge type. Details of calculation of these parameters can be found in the Methods section of Chapter 6.

A decision tree analysis works as follows: several questions regarding every data point represented by the multiple descriptive parameters are asked, one at a time. Depending on whether the answer to a question is true or false, the decision path bifurcates. The process is repeated until each of the paths reaches a ‘dead end’ node (known as the leaf) where there is only one possible answer to the question asked, no further bifurcations are possible. The number of times the questions were asked in a path becomes the depth of the decision tree. A part of the data with known outcomes is used for training the classification model, and another part is used for testing predictions. The decision tree we used was trained using DecisionTreeClassifier algorithm of scikit-learn[12] in Python.

7.3 Results and Discussion

The decision tree trained on 70% of the complete data set (**Methods section**) had a depth of 35 levels which had a prediction accuracy of 0.78 for the test set. Keeping in mind our goal that the model representation should fit an A4 size page, we restricted the decision tree to a maximum depth of 5. With this restriction of the depth to 5 levels, training with data between 40-70% did not significantly change the quality of output results. The overall accuracy for the test set was 0.71 and 0.73 respectively when trained with 50% or 70% of the data (Table 7.1). We finally selected a model that uses 70% of the data for training.

The simple decision tree model we developed is shown in Figure 7.1. Each node in this decision tree checks for a condition on one of the five descriptive physico-chemical parameters we chose for describing the amino acid mutations, and bifurcates towards the arrow if the condition is satisfied. The number of neutral and deleterious mutations at each node are also shown. At the last leaf of this decision tree, depending on the likelihood of having a neutral or deleterious mutation, a statistical conclusion about the neutrality is also made. It must be noted that a single mutation one is interested in may follow a decision path that ends at a leaf with higher prediction accuracies. The intent behind generating a decision tree that may be laid out on an A4 sheet was to facilitate a manual tracking of the logic of the decisions. A practical way one may use the decision tree shown in Figure 7.1 is for protein engineering, to check if deleterious mutations can be avoided. We validated that the predictions of deleterious mutations suggested by Figure 7.1 are non-trivial, and these effects can not be guessed with biochemical intuitions such as the distance from the active site. In deed, the deleterious mutations predicted from the tree were non-trivial. As an example, we illustrate in Figure 7.2, the fitness consequences of 4 substitutions in β-lactamase. The 5 physico-chemical properties corresponding to these mutations are given in Table 7.2. The decision flow for these mutations is shown in Figure 7.2. In the examples shown, there were 3 true predictions and one wrong prediction.

The decision tree developed was validated using the complete 2732 mutational effect scores from the deep mutational scan data of the protein β-glucosidase (Bgl3). The accuracy of prediction obtained was 0.75, comparable to the test set that had only variants of the proteins used for training, suggesting that the model may work for new data sets also. We compared our predictions with that of SNAP2 and found that the accuracy of SNAP2 predictions is lower for our training and test sets, the exact values being 0.62 and 0.61, respectively. For the Bgl3 data set the accuracy was 0.73 similar to the prediction quality we obtained with the decision tree developed.

While in nature it is believed that there is a unique relation between sequence, structure and function, there are wide gaps in understanding these relations. Predicting structure or function from sequence changes, is non-trivial, and it escapes simple rules and intuitions or codification of the knowledge or empiricism into formulas. However, with large data sets becoming available, it is clear that the many of the effects can be predicted reliably using the methods of AI, which use several descriptive parameters, complicated analyses, and computation. The philosophical debate that persists in the community of statisticians,[13] about whether the models should be predictive or interpretable thus becomes relevant for the protein mutational effects as well. Creating a decision tree based classification is one of the approaches used in AI. However in a typical AI decision tree approach, the emphasis is on developing the most detailed decision tree and to let the flow of decision logic happen in the computer. Since the focus of the AI methods was on effects prediction, we asked if an interpretation of the decision logic may be possible in the decision tree analysis of mutational effects. Without constraints, the depth of the decision tree depends upon the quality and size of the feature set as well as the data. To represent a legible decision tree in an A4 size, we worked with 5 levels of decision making. Practically the trade-off between the decision levels and the accuracy was not high (test set prediction accuracy of 0.73 with a depth of 5 compared to 0.78 for the complete tree with depth 35). While we understand the limitations of using a single decision tree such as its sensitivity to the training set and bias for the dominant class, we believe this hybrid approach fills the gap between the qualitative gut feelings about the factors governing the effects and the high accuracy models which remain black boxes in terms of understanding. We hope the transparent decision process helps develop intuitions about mutational effects and empowers the user to make educated guesses about the mutational effects. In Chapter 8 we make use of the newer developments in the field of explainable AI to develop interpretable as well as accurate model for mutational effect prediction.

7.4 Conclusions

Using artificial intelligence methods, we developed a simple decision tree which balances the simplicity of representation at the cost of a small loss of accuracy in predictions. The decision tree can be revisited in the future with more structured data. In a period where there is access to libraries of data from experiments and computations, we hope the simplicity of the approach serves as a reference for regaining intuitions about mutations.

Bibliography

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer, 2009.

[2] M. Hecht, Y. Bromberg, and B. Rost, “Better prediction of functional effects for sequence variants,” *BMC genomics*, vol. 16, no. 8, p. S1, 2015.

[3] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork,

A. S. Kondrashov, and S. R. Sunyaev, “A method and server for predicting damaging missense mutations,” *Nature Methods*, vol. 7, pp. 248–249, APR 2010.

[4] A. Niroula, S. Urolagin, and M. Vihinen, “Pon-p2: prediction method for fast and reliable identification of harmful variants,” *PloS one*, vol. 10, no. 2, p. e0117380, 2015.

[5] V. E. Gray, R. J. Hause, J. Luebeck, J. Shendure, and D. M. Fowler, “Quantitative missense variant effect prediction using large-scale mutagenesis data,” *Cell systems*, vol. 6, no. 1, pp. 116–124, 2018.

[6] M. A. Stiffler, D. R. Hekstra, and R. Ranganathan, “Evolvability as a Function of Purifying Selection in TEM-1 beta-Lactamase,” *Cell*, vol. 160, pp. 882–892, FEB 26 2015.

[7] A. Melnikov, P. Rogov, L. Wang, A. Gnirke, and T. S. Mikkelsen, “Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes,” *Nucleic Acids Research*, vol. 42, no. 14, 2014.

[8] P. Mishra, J. M. Flynn, T. N. Starr, and D. N. A. Bolon, “Systematic Mutant Analyses Elucidate General and Client-Specific Aspects of Hsp90 Function,” *Cell Reports*, vol. 15, pp. 588–598, APR 19 2016.

[9] L. Brenan, A. Andreev, O. Cohen, S. Pantel, A. Kamburov, D. Cacchiarelli, N. S. Persky, C. Zhu, M. Bagul, E. M. Goetz, A. B. Burgin, L. A. Garraway, G. Getz, T. S. Mikkelsen, F. Piccioni, D. E. Root, and C. M. Johannessen, “Phenotypic Characterization of a Comprehensive Set of MAPK1/ERK2 Missense Mutants,” *Cell Reports*, vol. 17, pp. 1171–1183, OCT 18 2016.

[10] J. Weile, S. Sun, A. G. Cote, J. Knapp, M. Verby, J. C. Mellor, Y. Wu, C. Pons, C. Wong, N. van Lieshout, F. Yang, M. Tasan, G. Tan, S. Yang, D. M. Fowler, R. Nussbaum, J. D. Bloom, M. Vidal, D. E. Hill, P. Aloy, and F. P. Roth, “A framework for exhaustively mapping functional missense variants,” *Molecular Systems Biology*, vol. 13, DEC 2017.

[11] S. Henikoff and J. Henikoff, “Amino-acid substitution matrices from protein blocks,” *Proceedings of the National Academy of Sciences, USA*, vol. 89, pp. 10915–10919, NOV 15 1992.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, “Scikit-learn: Machine learning

in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[13] L. Breiman, "Statistical modeling: The two cultures," *Statistical Science*, vol. 16, p. 199, 2001.

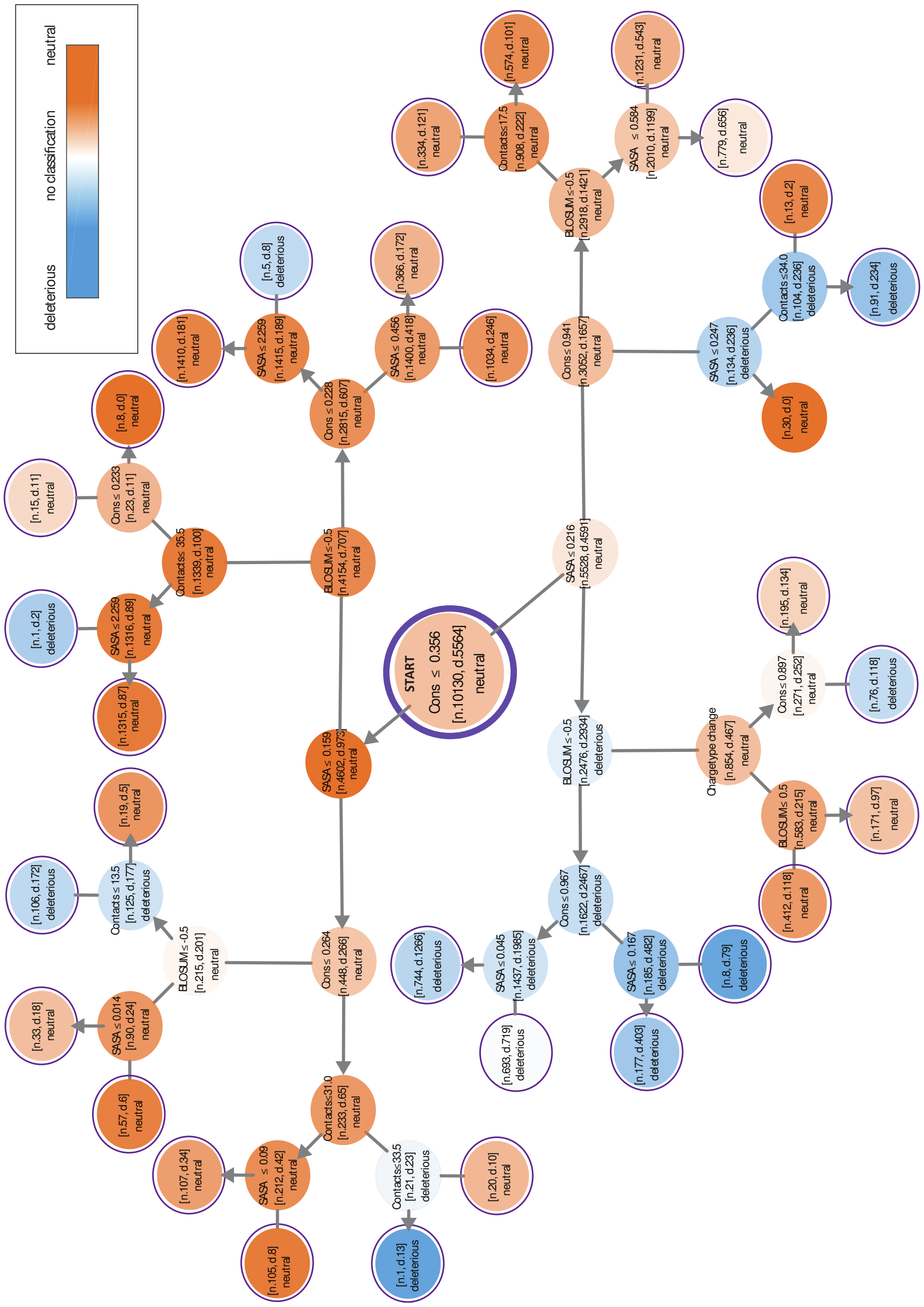
deleterious
no classification
neutral
START
Cons ≤ 0.356
[n.10130, d.5564]
neutral
SASA ≤ 0.159 [n.4602, d.973] neutral
BLOSUM ≤ -0.5 [n.4154, d.707] neutral
Cons ≤ 0.228 [n.2815, d.607] neutral
SASA ≤ 2.259 [n.1415, d.189] neutral
[n.1410, d.181] neutral
[n.5, d.8] deleterious
SASA ≤ 0.456 [n.1400, d.418] neutral
[n.366, d.172] neutral
[n.1034, d.246] neutral
Contacts ≤ 35.5 [n.1339, d.100] neutral
Cons ≤ 0.233 [n.23, d.11] neutral
[n.8, d.0] neutral
[n.15, d.11] neutral
SASA ≤ 2.259 [n.1316, d.89] neutral
[n.1, d.2] deleterious
[n.1315, d.87] neutral
Cons ≤ 0.264 [n.448, d.266] neutral
BLOSUM ≤ -0.5 [n.215, d.201] neutral
Contacts ≤ 13.5 [n.125, d.177] deleterious
[n.19, d.5] neutral
[n.106, d.172] deleterious
SASA ≤ 0.014 [n.90, d.24] neutral
[n.33, d.18] neutral
[n.57, d.6] neutral
Contacts ≤ 31.0 [n.233, d.65] neutral
SASA ≤ 0.09 [n.212, d.42] neutral
[n.107, d.34] neutral
[n.105, d.8] neutral
Contacts ≤ 33.5 [n.21, d.23] deleterious
[n.20, d.10] neutral
[n.1, d.13] deleterious
SASA ≤ 0.216 [n.5528, d.4591] neutral
Cons ≤ 0.941 [n.3052, d.1657] neutral
BLOSUM ≤ -0.5 [n.2918, d.1421] neutral
Contacts ≤ 17.5 [n.908, d.222] neutral
[n.574, d.101] neutral
[n.334, d.121] neutral
SASA ≤ 0.584 [n.2010, d.1199] neutral
[n.1231, d.543] neutral
[n.779, d.656] neutral
SASA ≤ 0.247 [n.134, d.236] deleterious
Contacts ≤ 34.0 [n.104, d.236] deleterious
[n.13, d.2] neutral
[n.91, d.234] deleterious
[n.30, d.0] neutral
BLOSUM ≤ -0.5 [n.2476, d.2934] deleterious
Chargetype change [n.854, d.467] neutral
Cons ≤ 0.897 [n.271, d.252] neutral
[n.195, d.134] neutral
[n.76, d.118] deleterious
BLOSUM ≤ 0.5 [n.583, d.215] neutral
[n.171, d.97] neutral
[n.412, d.118] neutral
Cons ≤ 0.967 [n.1622, d.2467] deleterious
SASA ≤ 0.045 [n.1437, d.1985] deleterious
[n.744, d.1266] deleterious
[n.693, d.719] deleterious
SASA ≤ 0.167 [n.185, d.482] deleterious
[n.8, d.79] deleterious
[n.177, d.403] deleterious

Figure 7.1: Decision tree for mutational effects. A decision tree of maximum depth 5 levels was developed. The decision flow begins with the central node which explores whether conservation is less than 0.356 as the first question. At every node, if the condition is satisfied the decision flows along the arrow and along the straight line otherwise. A notation such as (n.212, d.42) indicates that in the training data 212 neutral and 42 deleterious mutations reached this node. The colorbar on the top is a qualitative illustration of the color code for the decision nodes, depending on whether the node represents a deleterious, or neutral mutation or can not be classified as one of these.

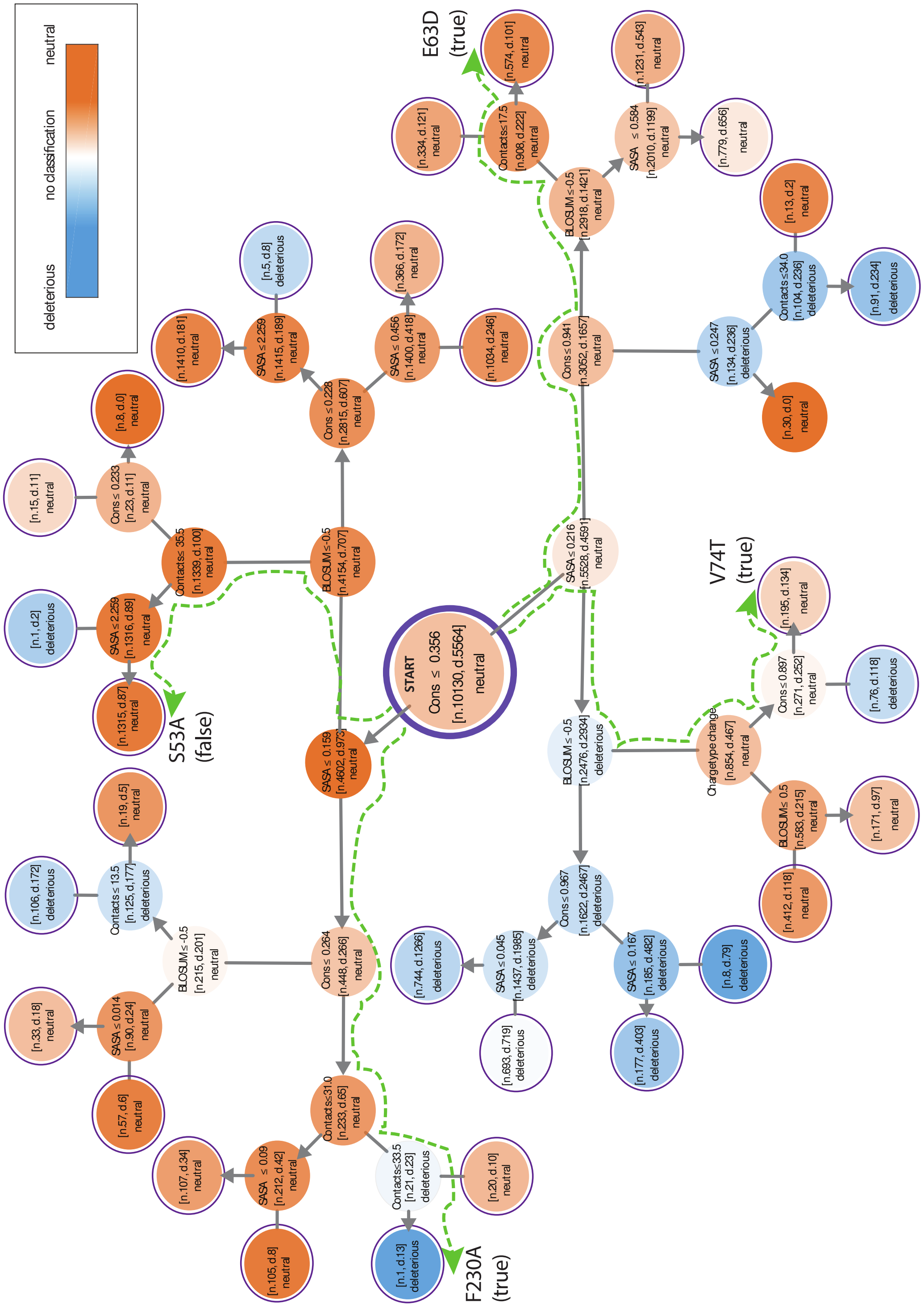

Figure 7.2: Illustration of decision tree usage. The effects of four mutations (S53A, E63D, V74T, F230A) is illustrated in the decision tree. The parameters used for this analysis are given in Table 7.2. Green lines show the paths along the decision tree for these mutations. Three of them resulted in true predictions, while the prediction of S53A using this tree was not correct.

Data used for training	Test/ Training	True Neutral (TN)	False neutral (FN)	True deleterious (TD)	False deleterious (FD)	Accuracy
70%	Training	8228	2550	3014	1902	0.72
70%	Test	3551	1055	1357	764	0.73
50%	Training	5668	1615	2362	1565	0.72
50%	Test	5672	1659	2340	1540	0.71

Table 7.1: The contingency table for the training and test sets when 70% and 50% of the data was used for training. Accuracy was defined as the ratio of number of true predictions (TN+TD) to the total number of data points.

Variant	Conservation	No. of contacts	SASA (nm^2)	BLOSUM substitution score	Charge type change
S53A	0.286	8	0.941	1	Yes
E63D	0.487	4	1.308	2	No
V74T	0.408	15	0.000	0	Yes
F230A	0.179	32	0.117	-2	No

Table 7.2: The physico-chemical parameters used for performing a trial of the decisions for the 4 mutations from β-lactamase shown in Figure 7.2 are given here.

Chapter 8

Interpreting Mutational Effects Predictions, One Substitution at a Time

Abstract

Artificial intelligence (AI) based methods for mutational effects predictions are improving in accuracy, because of the exhaustive experimental data they are trained on, and advances in algorithms. As the prediction quality improves, the next natural question to ask countering the 'black box' image of AI is if the predictions can be interpreted. We applied one of the approaches developed in the field of explainable AI, to decipher the factors contributing to the changes in cellular fitness and protein solubility arising from mutations. Juxtaposing the observations from these two desirable outcomes and focusing on the individual factors uncovers the contributions and quantifies the intuitions about how different factors such as conservation, distance from the catalytic site affect fitness and solubility. Embedding interpretability along with the prediction algorithms will enable transparency and inspire confidence into models as well as contribute to the understanding of how mutations affect proteins.

8.1 Introduction

Ability to make reliable mutational effects predictions using machine learning approaches brings one to a natural point of asking the questions that are being asked in other areas where machine learning is used, such as why the predictions should be trusted[1] or alternatively if the predictions can be interpreted[2, 3]. The philosophical debate about predictability versus interpretability[4] is being revisited in several areas of machine learning, specifically when one is interested in controlling the effects by developing an understanding for the contributing factors. In this chapter this notion of interpretability or explainability in the context of mutational effects prediction is introduced. Interpretability begins with a small shift in perspective, from asking how do various factors contribute to the set of predictions, to what is the contribution of the various factors to an individual prediction. In a linear regression model, knowing the measured outcome, it is trivial to understand the relative contributions from the different factors. The same is not true while working with machine learning models, which have non-trivial and non-explicit relations between the inputs and the outcome. The simple decision tree developed in the previous chapter was a step towards that, but the interpretability was gained at the cost of accuracy of predictions.

We use the method SHAP (SHapley Additive exPlanations)[5, 6] that is being used in several areas of machine learning, to interpret the contributions to the mutational effects. SHAP is based on the game theoretical questions raised by Shapley[7] about how the gain can be shared by different contributing players. In SHAP, the feature contributions are additive, thus making their relation to the outcome easy to interpret. We apply SHAP to interpret the outcomes of fitness[8] and solubility[9] in the deep mutational scans of β-lactamase protein. The interpretability allows us to revisit the classical intuitions on how different factors can influence mutations from a quantitative perspective.

8.2 Methods

Deep mutational scan data. The deep mutational scan data of the mutational effects of β-lactamase on fitness was obtained from Stiffler et al. 2015[8] and solubility from Klesmith et al. 2016[9]. We use the yeast surface display (YSD) data on the effects of mutations on solubility [20] and the changes in relative fitness of E. coli when a mutant containing strain is challenged with 2500 $\mu g/ml$ ampicillin [19]. The analyses on the solubility were presented for the substitutions on positions 61-215 [20], with protein data bank identity 1M40,[10] and for consistency, we chose to work with the same set of mutations both for the solubility and cellular fitness.

Descriptive variables for AI model. We used distance from the catalytic site in addition to the 17 features that were used to develop Deep2Full[11] in Chapter 5. Distance from catalytic site was calculated as the distance of the amino acid from the nearest catalytic site. Calculation of other variables can be found in the **Methods** section of Chapter 5.

All the 18 variables are referred as follows in figures: Solvent accessible surface area - *SASA*, Secondary structure (Ordered/disordered) - *SS*, Contacts - *Contacts*, Distance from the catalytic site - *Catalytic_ dist*, Average commute time - *Av_ commutetime*, BLOSUM62 substitution matrix score - *Blosum*, Hydrophobicity of the wild type amino acid - *Wt_hb*, Hydrophobicity of the amino acid after mutation - *Mut_hb*, Position specific scoring matrix score for the wild type amino acid - *PSSM_w*, Position specific scoring matrix score for the amino acid that is substituted by - *PSSM_m*, Conservation - *Conservation*, Average co-evolutionary correlation - *Av_ corr*, Degree centrality - *Degree*, Betweenness centrality - *Betweenness*, Closeness centrality - *Closeness*, Eigenvector centrality - *Eigenvector*, Impact - *Impact* and Dependency - *Dependency*

AI model. Using the 18 descriptive parameters and the experimental measurements for each of the mutations, the AI analyses were performed using Python. For predicting the effects, we used XGRegressor implemented in the XGBoost package. 75% of the mutational data from amino acids 61-215 was used for training and the remaining 25% for predictions. As shown in Figure F.1 of Appendix F, the Pearson correlation coefficients for the test sets compared to the experiments were good (0.88 for fitness and 0.79 for solubility).

Interpretable AI model. SHapley Additive explanation (SHAP) uses the formalism where an explanation model g is defined in terms of the parameter set z_i' defining each instance (in our case each individual mutation) and their corresponding additive contribution ϕ_i weights.[5, 6]

$g(z') = \phi_0 + \Sigma_i\ \phi_i z_i'$

The explanatory model is subject to three conditions known as:

Local accuracy – which ensures that it matches the calculated effect f(z) when $z' = z$, i.e., $g(z') = f(z)$ when $z' = z$,

Missingness – which ensures that if a variable $z' = 0$, then the weight corresponding to it, $\phi_i = 0$.

Consistency – when an input's contribution increases or stays the same regardless of the other inputs, then its weight should not decrease.

By solving for these three conditions, one obtains the SHAP contribution weights corresponding to each individual input instance. We used the SHAP implementation by Lundberg (https://github.com/slundberg/shap) for performing the interpretable AI calculations, where corresponding to each mutational effect calculation, all the are determined. The results presented in this work discuss these SHAP weight factors

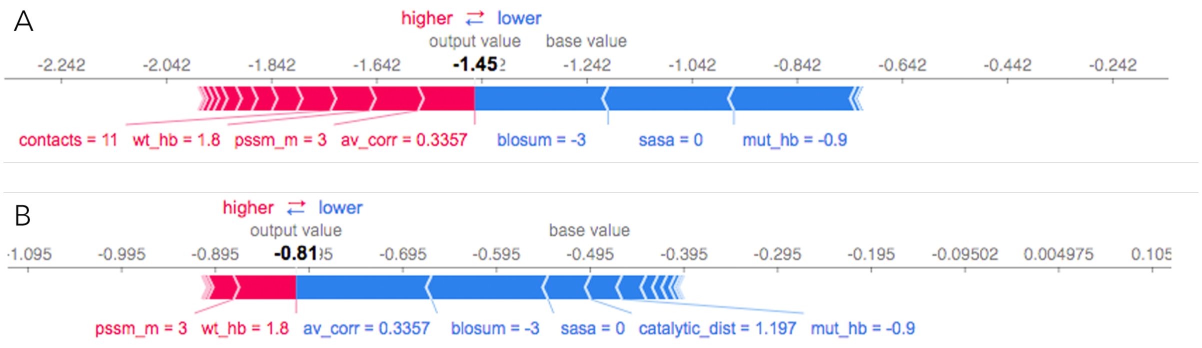

Figure 8.1: Decomposing the contributions. Illustration of the contributions of various factors to the effects of mutating Alanine in position 79 to Tyrosine (A79W) in β-lactamase: A. Fitness effect B. Solubility effect. As indicated by the direction of the arrows, the factors in pink contribute to an increase in the fitness or solubility and those in blue have the opposite effect. Whether a specific factor tends to increase or decrease the mutational effect depends on the individual case. The descriptive parameters are labeled along with the values they assume in this specific instance, for the specific mutation. The illustrations are generated using the Python implementation of SHAP ((https://github.com/slundberg/shap)).

8.3 Results and Discussion

8.3.1 Noting the contributions from the individual factors to individual mutations.

The XGRegressor model we used for making the predictions of the fitness and solubility gave good performance on a statistical level (Pearson correlation coefficients of 0.88, 0.79 for fitness and solubility, F.1 of Appendix F). Taking confidence in these predictions, we applied SHAP method to mutational effects calculations, and obtained the contributing factors in each individual mutation. Figure 8.1 illustrates the predictions for a specific mutation A79W and the factors contributing to it. The predicted fitness (-1.45) and solubility (-0.81) for this mutation compare well with the experimental observations (-1.64, -1.10 respectively). The interpretable aspect of the prediction is shown in the decomposition of the various factors, firstly segregated by positive and negative contributions: factors labeled in pink aiding a better fitness or solubility, and those in blue having the opposite effect. The length of the bar representing each factor reflects the magnitude of its contribution to the specific outcome. For example, the statistical descriptor of the likelihood of substitution (BLOSUM62) which has a value of -3 contributes in a comparable way to reducing both the fitness and the solubility. On the other hand, the average co-evolutionary relation an amino acid shares with all other amino acids (avg_corr) which has a value of 0.3357 has opposite effects on fitness and solubility.

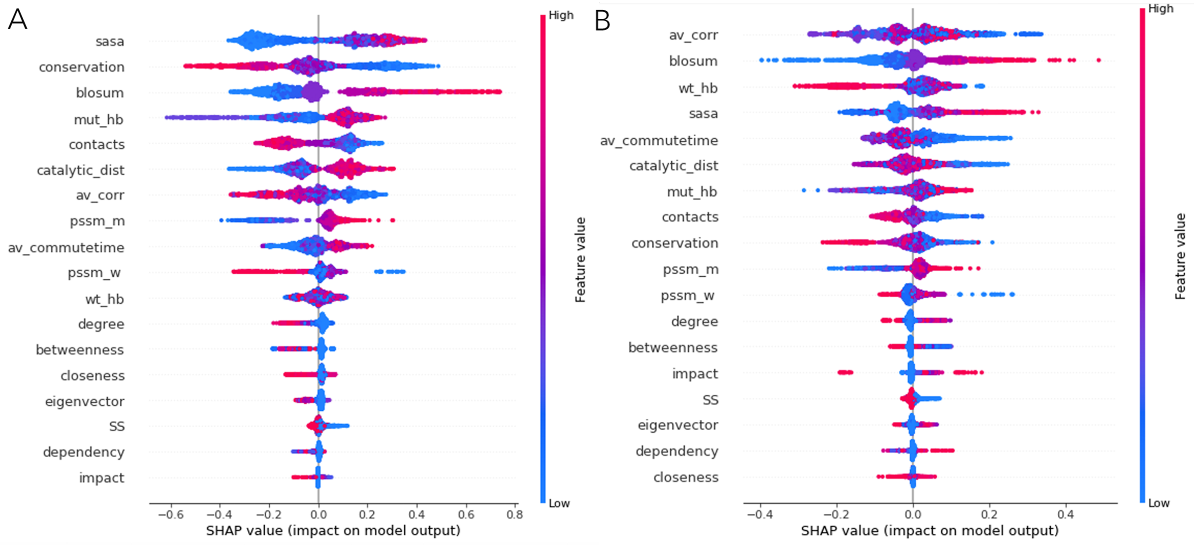

Figure 8.2: Summarizing the contributions. We analyzed the deep mutational scan data where the consequences of any of the 19 possible amino acid substitutions at each of the positions (61-215) were measured. The individual contributions to A. fitness and B. solubility obtained from each of these mutations are summarized in the plot. Along each line, one finds the name of the descriptive parameter, a distribution of the SHAP values across the complete set of mutations, along with the color indicator of the fitness/solubility outcome associated with mutation.

8.3.2 Summarizing the contribution of the individual factors in the complete set

The impact obtained from individual factors is then summarized to understand the variables that have the most significant role in the set of predictions (Figure 8.2). Observing the range of the values assumed, one can infer that conservation, SASA, BLOSUM, along with the hydrophobicities of the wild type and mutant amino acids contribute significantly to both solubility and fitness. However, one can notice in the illustrations that at times the distribution of the data points labeled pink (higher outcome) and blue (lower outcome) for the some of the variables is different when comparing fitness and solubility (Figures 8.2A and 8.2B).

8.3.3 Identifying factors making correlated contributions to fitness and solubility.

We further examined the contributions of each of the parameters to fitness and solubility in the same analysis, to see if they are correlated, anti correlated or uncorrelated. As shown in Figure 8.3, the contributions from three variables, conservation, BLOSUM and number of contacts have a good positive correlation. The contributions of other

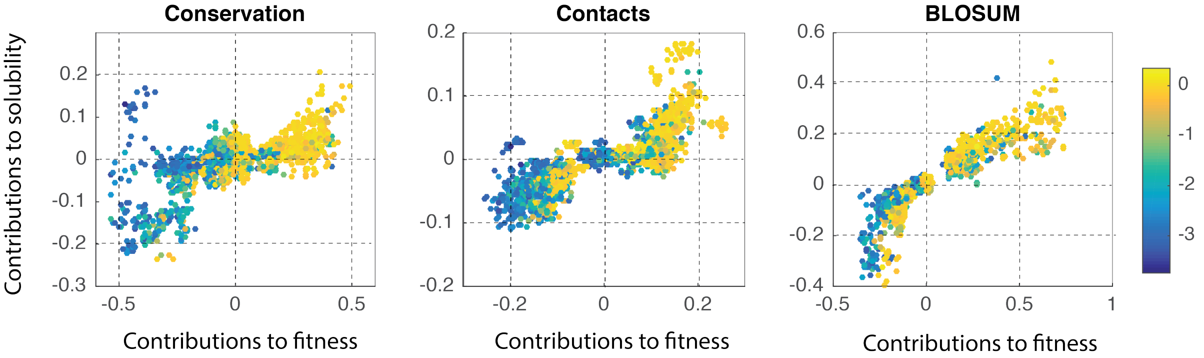

Figure 8.3: Correlation of contributions to solubility and fitness. The SHAP values defining the contribution of each variable to fitness and solubility are shown. From the data it is clear that the contributions from these three descriptive parameters, conservation, number of contacts and BLOSUM are mostly correlated i.e., if the parameter contributes to an increase in fitness it also contributes to an increase in the solubility. This is not true for all variables, which are shown in the Figures F.2 to F.4 of Appendix F, where they are uncorrelated or negatively correlated. The color bar represents the experimentally observed fitness changes among all the mutations studied.

variables to the fitness and solubility do not have strong correlations (Figures F.2 and F.3 of Appendix F).

8.3.4 Extracting intuitive patterns about the effect of different factors.

There are several intuitions about how factors such as conservation or distance from the catalytic site should influence the protein function or its solubility. Many of these intuitions still need to be quantified. Klesmith et al.[9] performed a very careful analysis using naÃŕve Bayesian classification to clarify the chance that a mutation characterized by a parameter is statistically likely to be deleterious or neutral. They observe an interesting trade-off such as that the distance from the catalytic site has opposite effects on solubility and fitness. We ask if one can go beyond the classification models to quantify these dependencies using SHAP analysis. The scatter plots in Figures 8.4A, 8.4B show that the relation of fitness and solubility to conservation is not easy to infer. Naively, while observing the relation of the outcomes to a single variable in a multi-factorial system, one may expect either a poor correlation or even a lack of it. However, when we plot the contributions to fitness and solubility using the SHAP analysis (Figures 8.4E, 8.4F), the dependence of the component becomes much more predictable. Figures 8.4C, 8.4D, 8.4G, 8.4H illustrates how the distance from the catalytic effect makes predictable contributions to the fitness, and solubility. Interestingly the two outcomes (Figures 8.4G, 8.4H) show an opposite dependence on the distance from the catalytic site, as was seen using a classification model [9]. The lower panel of Figure 8.4 shows a detailed quantitative relation between the contributing factors and the measurable outcomes. Similarly, Figure F.4 of Appendix F illustrates the effect of the number of contacts.

8.3.5 From non-linearity to linearity.

The above analysis provides two new perspectives. Firstly it uncovers the patterns in the contributions from the individual factors, when they exist (such as in Figures 8.4E-H) and secondly the additive nature of the SHAP contributions makes it possible to obtain an outcome in relatively simple terms, unlike with the machine learning algorithms. It is true that the non-linearity is only masked by making suitable transformations. Thus the interpretability converts the non-linear predictive models into interpretable, simplified representation, in which one can audit the contributions from the different factors to, for example, the simultaneous and conflicting requirements on solubility and fitness.

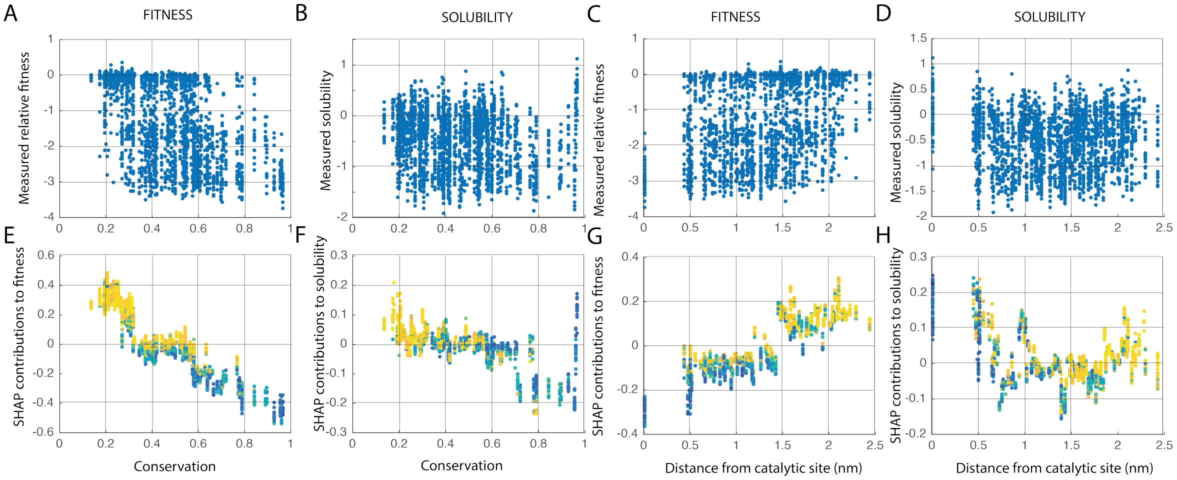

Figure 8.4: Extracting relations. The scatter plots of A-D of fitness and solubility relative to conservation and the distance from the catalytic site do not show a clear pattern of what one can expect from substituting an amino acid with high conservation or further away from catalytic site. On the contrary, E,F. the SHAP contributions highlight a very clear pattern of reducing SHAP values with increasing conservation which suggests that the fitness and solubility decrease with the substitution of a conserved amino acid. G, H. SHAP contributions show a contrasting behavior where the distance from the catalytic site has opposite effects on fitness and solubility, in line with the classifications[9]. The colorbar is the same as in Figure 8.3, and represents the observed fitness changes.

8.3.6 Perspective.

Seeing the emphasis and developments on explainability of AI predictions in several areas of science and engineering, it is clear that the mutational effects predictions should also benefit from such analyses. The explainability analyses can serve several purposes:

Validation of Correctness – An important consequence of explaining the effects is that by validating that the factors that one believes are indeed the most relevant ones in calculations, a sense of correctness of individual predictions may be developed.
Protein Engineering – The analyses such as understanding the trade-offs between solubility and fitness[9] have been strongly motivated from the perspective of designing better proteins. The same is true when a direct correlation between measured fitness and solubility cannot be inferred (Figure F.5 of Appendix F), where the individual components are more predictable and hence reliable.
Developing intuitions – The developments in the fields of deep mutational scan have a great value to add to the intuitions that are pedagogically taught, such as larger conservation implies greater impact. The patterns of dependence of the mutational effects on the individual parameters allows one to go beyond predictions to learning and developing of rules.
Understanding protein function is not easy. At times a single mutation can lead to deleterious effects, and yet evolutionarily one sees homologous proteins with as little as 50% sequence identity performing similar functions. One has to resort to advanced AI methods to predict the effects of one or more mutations, to note how mutations affect, or compensate for each other, or to reduce the experimentation required. Introducing explainability into the analyses can potentially help in an improved learning about how mutations affect the proteins.

8.4 Conclusions

Artificial intelligence based models are making reliable predictions of the mutational effects, whether it is changes in solubility or the cellular fitness. In this work we asked the next natural question which is, if there is access to a large pool of systematic mutational scans, and reliable AI based models, can one explain the different factors that contribute to each individual mutational effect? In asking so, with the standard tools that are available, we uncover quantitative patterns in contributions of the different variables to fitness and solubility, sometimes inline and at times opposing.

Bibliography

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, “" why should i trust you?" explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[2] A. Datta, S. Sen, and Y. Zick, “Algorithmic transparency via quantitative input

influence: Theory and experiments with learning systems," in *2016 IEEE symposium on security and privacy (SP)*, pp. 598–617, IEEE, 2016.

[3] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and information systems*, vol. 41, no. 3, pp. 647–665, 2014.

[4] L. Breiman *et al.*, "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," *Statistical science*, vol. 16, no. 3, pp. 199–231, 2001.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, pp. 4765–4774, 2017.

[6] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:1802.03888*, 2018.

[7] L. S. Shapley, "A value for n-person games, volume ii of contributions to the theory of games," 1953.

[8] M. A. Stiffler, D. R. Hekstra, and R. Ranganathan, "Evolvability as a Function of Purifying Selection in TEM-1 beta-Lactamase," *Cell*, vol. 160, pp. 882–892, FEB 26 2015.

[9] J. R. Klesmith, J.-P. Bacik, E. E. Wrenbeck, R. Michalczyk, and T. A. Whitehead, "Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning," *Proceedings of the National Academy of Sciences*, vol. 114, no. 9, pp. 2265–2270, 2017.

[10] G. Minasov, X. Wang, and B. K. Shoichet, "An ultrahigh resolution structure of tem-1 β-lactamase suggests a role for glu166 as the general base in acylation," *Journal of the American Chemical Society*, vol. 124, no. 19, pp. 5333–5340, 2002.

[11] C. Sruthi and M. Prakash, "Deep2full: Evaluating strategies for selecting the minimal mutational experiments for optimal computational predictions of deep mutational scan outcomes.," *PloS one*, vol. 15, no. 1, 2020.

Chapter 9

Using Deep Mutational Scan Data for Understanding Site Specific Codon Usage Bias

Abstract

Codon usage bias defines the unequal use of the multiple codons which encode the same amino acid. This preference which varies across genomes, and even across genes has been interpreted using arguments on the efficiency of translation, and reduction of the errors due to mutations. The latter has been mostly calculated using generalized substitution matrices and reduction in chemical distance across amino acid substitutions. In this work, we use deep mutational scan data to ask if the site-specific preference for the codons can be predicted. Interestingly, the sites for which the codon predictions are correct are complementary to those from tRNA availability, and comparable in number.

9.1 Introduction

Amino acids are encoded by three-nucleotide codes, known as the codons. Other than tryptophan and methionine, all other amino acids are encoded by multiple codons, represented by different permutations of the nucleotides. Interestingly these synonymous codons do not occur with the same frequency, and the preference of some synonymous codons over others in the DNA is known as the codon usage bias.[1, 2] There is a strong correlation between GC content of the genome and the codon bias. However, the preferred codons used for encoding an amino acid varies across genomes[1] as well as across the genes within the same genome.[3] The observed bias in codon usage has been interpreted using the mutation-selection-drift balance model[4–6] which incorporates two major factors that contribute to codon usage bias, the mutational bias and the selection force on synonymous codons. Mutational bias refers to the variations in the probabilities of different codons getting mutated. These variations can give rise to non-uniform frequency of codons. Since mutational biases are organism specific, the codon usage also varies between organisms.[7]

Selection for translational efficiency that act on synonymous codons is another explanation for the codon usage bias.[8] The correlation between codon usage bias and gene expression level[9] supports this hypothesis.[3] The observation that the availability of tRNA with the corresponding anti-codon for preferred codons is higher[3, 8, 10] suggests the possibility of more accurate and faster translation of these codons resulting in higher translational efficiency and a positive selection for these codons. Apart from the tRNA availability, several other factors, such as the co-occurrence bias or codon pair-bias, are believed to determine the speed of translation. Co-occurrence bias which is the clustering of synonymous codons corresponding to the same tRNA in parts of the gene, facilitates the availability of tRNA near ribosome faster by recharging the tRNA that exited the ribosome and hence improve the translation efficiency.[11] Codon pair-bias, which is a result of preference for certain pairs of codons over others, also affects the choice of codon.[12, 13] In addition to translational efficiency, the codon usage is also determined by its effect on the transcription,[14] the stability of mRNA structure[15] and half life[16] and its influence on co-translational protein folding.[17]

All the above mentioned selection pressures act on synonymous codons. But there have been theoretical models which showed that selection at the amino acid level to reduce the deleterious effects of substitutions also affect the codon usage.[18, 19] In these models, the amino acid substitution effects were determined based on substitution matrices which are not site-specific such as Grantham matrix .[18] In fact, the selection pressure on the amino acids is so important that mutation-selection arguments have been used to understand the overall architecture of the genetic code.[20–23] It has been postulated that the evolution of codon-amino acid pairs has been in such a way as to reduce the effect of nucleotide substitution, defined as the chemical distance between amino acids.[20–23]

Now with the availability of the phenotypic fitness effects data of all single amino acid substitutions from deep mutational scans, instead of general amino acid substitution matrices based on the chemical distances, a fitness based selection criterion can be applied. A deep mutational scan study of TEM-1 β-lactamase focussed on the measurements of the fitness effects of mutations, and its bearings on the understanding of the general codon architecture. The significant effect of synonymous codons on the fitness was mainly seen for amino acids 2 to 10 of TEM-1 β-lactamase.[24] By averaging the contributions of the synonymous and non-synonymous codons resulting from 1-, 2-, and 3-base pair substitutions, they could show a general trend of declining fitness with higher number of nucleotide substitutions.[24] In this work, we use deep mutational scan data of all single amino acid substitutions to ask a different question, if the site specific preferences for the codon usage bias can be predicted.

9.2 Results and Discussion

9.2.1 Average deleteriousness upon single nucleotide substitution

The wild type amino acid at each position was used as a reference, and all codons that could in principle code for this amino acid are compared to see if an ideal codon could be identified. The comparison was based on the deleteriousness that could arise by starting from a choice of the codon, and tolerating up to one single nucleotide substitution. This tolerance leads to 9 possible substitutions, some of which are synonymous, and others are non-synonymous including stop codons. The relative-fitness measurements for the non-synonymous amino acid substitutions could be obtained from deep mutational scan experiments, such as the ones for TEM-1 β-lactamase[25] and APH(3′)-II[26]. These experiments report the mutational effect score for a substitution of amino acid a to b at site i is the relative-fitness,[25] $R_i^{a,b} = log_{10}\left(f^{b,i}/f^{a,i}\right)$ where f is the ratio of allele counts in the selected and unselected population was used as a measure of phenotypic outcome of mutations. While considering the fitness effect upon mutation, it is also important to account for the translational efficiency that comes from the availability of the codon after substitution, so that the mutated mRNA sequence can be translated. For every amino acid site in the protein, the averaged deleterious consequence of one single substitution was calculated using the relative-fitness measurements for these 9 substitutions from the deep mutational scan data: $\sum_{b=1}^{9} tRNA\, R_i^{a,b}/9$. For the stop codons, the maximum loss of fitness ($R_i^{a,\text{stop}} = -4$) seen in these data sets was assigned. The gene copy numbers in *E. coli* obtained from http://gtrnadb.ucsc.edu were used as a surrogate for the $tRNA$ availability. For APH(3′)-II the missing variant effects were predicted using the neural network model developed by training on the data of other variants as has been done in Chapter 5. The average deleteriousness consequence of a single mutation was calculated for every possible codon that can encode the wild type amino acid at that site. It is apparent from Figures 9.1 and 9.2 and Figures G.1 and G.2 of Appendix G that the codons can have very significant differences relative to one another in terms of stability against potential loss of fitness. We explored to see whether this loss of fitness for the individual amino acids was correlated with the codon usage bias

(Figure 9.3 and Figure G.3 of Appendix G). Relative codon usage for a codon calculated as the ratio of observed frequency of that codon in the complete genome to the frequency expected if all synonymous codons were used equally for E. coli was obtained from Sharp et al.[27] No specific pattern was observed, however we worked with the minimization of this potential loss of fitness in the section below.

9.2.2 Complementarity of tRNA availability and minimization of loss of fitness criteria

We used the minimization of the deleteriousness calculated as above, as a criterion to choose the ideal codon. Clearly, this approach was not suited for all the amino acid positions, but only for 89 out of the 250 amino acids in β-lactamase and 83 out of the 253 positions of APH(3′)-II. Interestingly, the codons that are identified as ideal by the availability of tRNA are comparable in number (75 in β-lactamase and 74 in APH(3′)-II) and are complementary to the loss of fitness criterion (Figure 9.4).

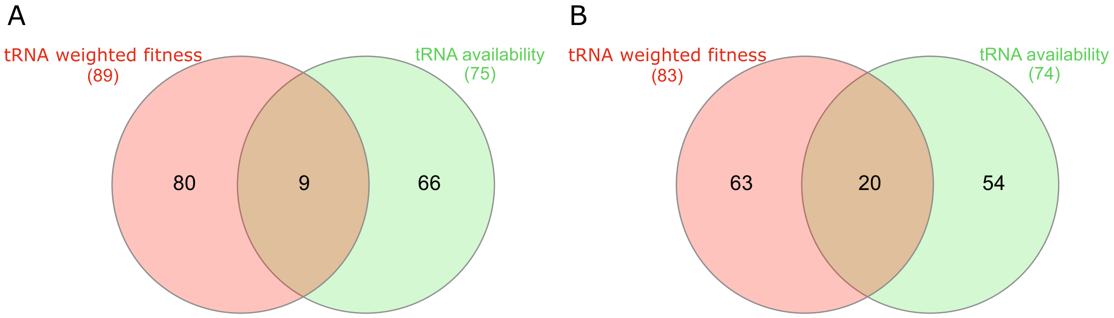

Figure 9.4: Number of correctly predicted synonymous codons based on average of tRNA availability weighted fitness and tRNA availability are compared. The number of correct predictions in each case and the overlap between them are shown for (A) β-lactamase and (B) APH(3′)-II.

9.2.3 Search for patterns

While the complementarity between tRNA availability and the loss of fitness criteria was encouraging, we analyzed the data to see if there are any patterns in the suitability of one criterion or the other. It is generally observed that major codons are chosen for functionally important amino acids[28, 29] as well as secondary structure and codon choices are related.[30] We investigated whether of the two complementary predictors, loss of fitness or tRNA availability, one of them better predicts the choice of codons for an amino acid in a given secondary structure. Both the criteria had comparable accuracy for all secondary structures (Tables G.1 and G.2). Similar estimation of accuracy of prediction was performed for the different amino acid depending on their chemical nature, or conservation, and no significant differences were found. At this point, although the

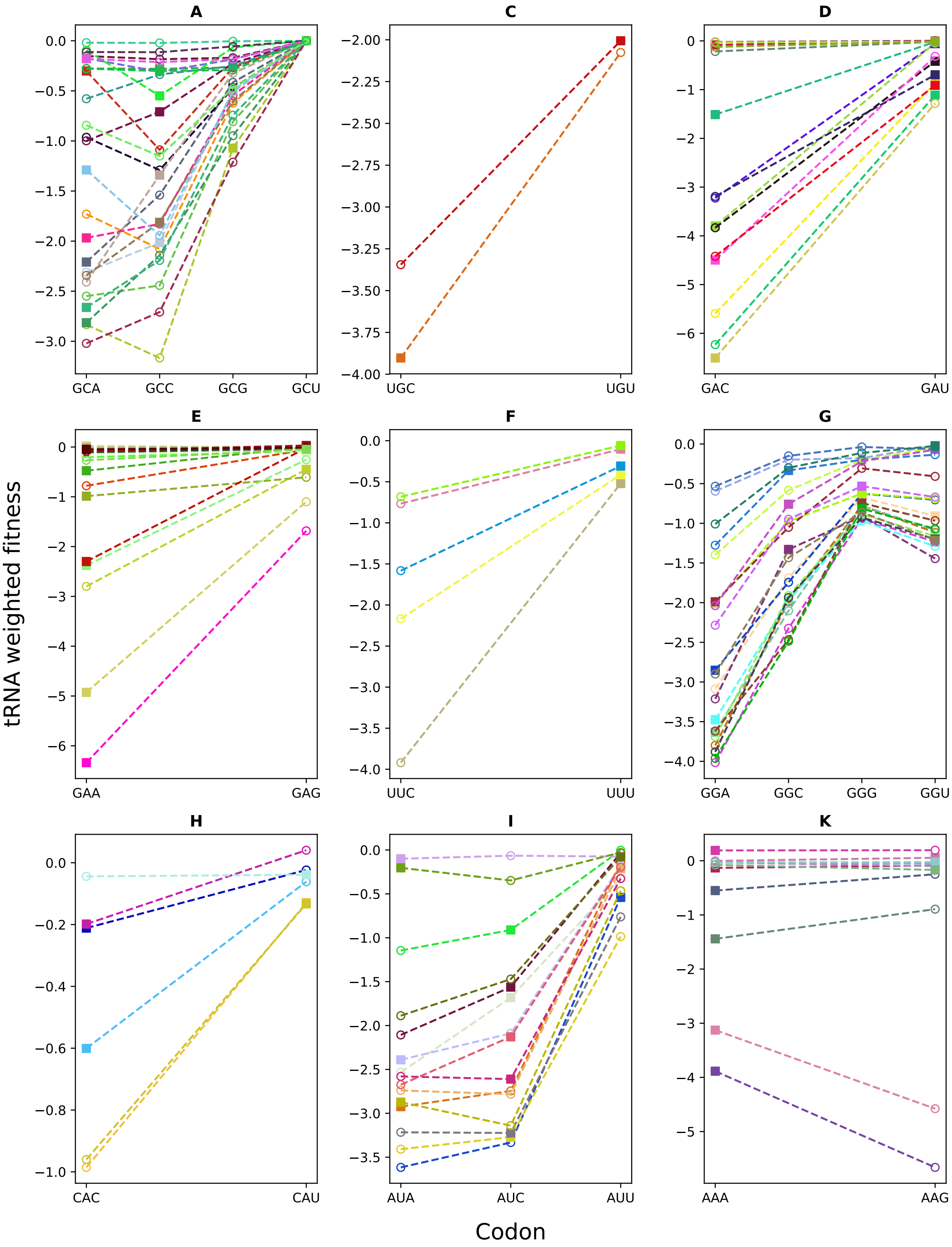

Figure 9.1: The average of tRNA weighted fitness scores of variants that are possible with one nucleotide change for every synonymous codon of amino acids except M and W in β-lactamase are shown. Colours within each subplot correspond to different residue positions in the protein with the specific amino acid. Square symbol indicates the codon in the WT sequence. The results for the remaining 9 amino acids are given in Figure 9.2
.

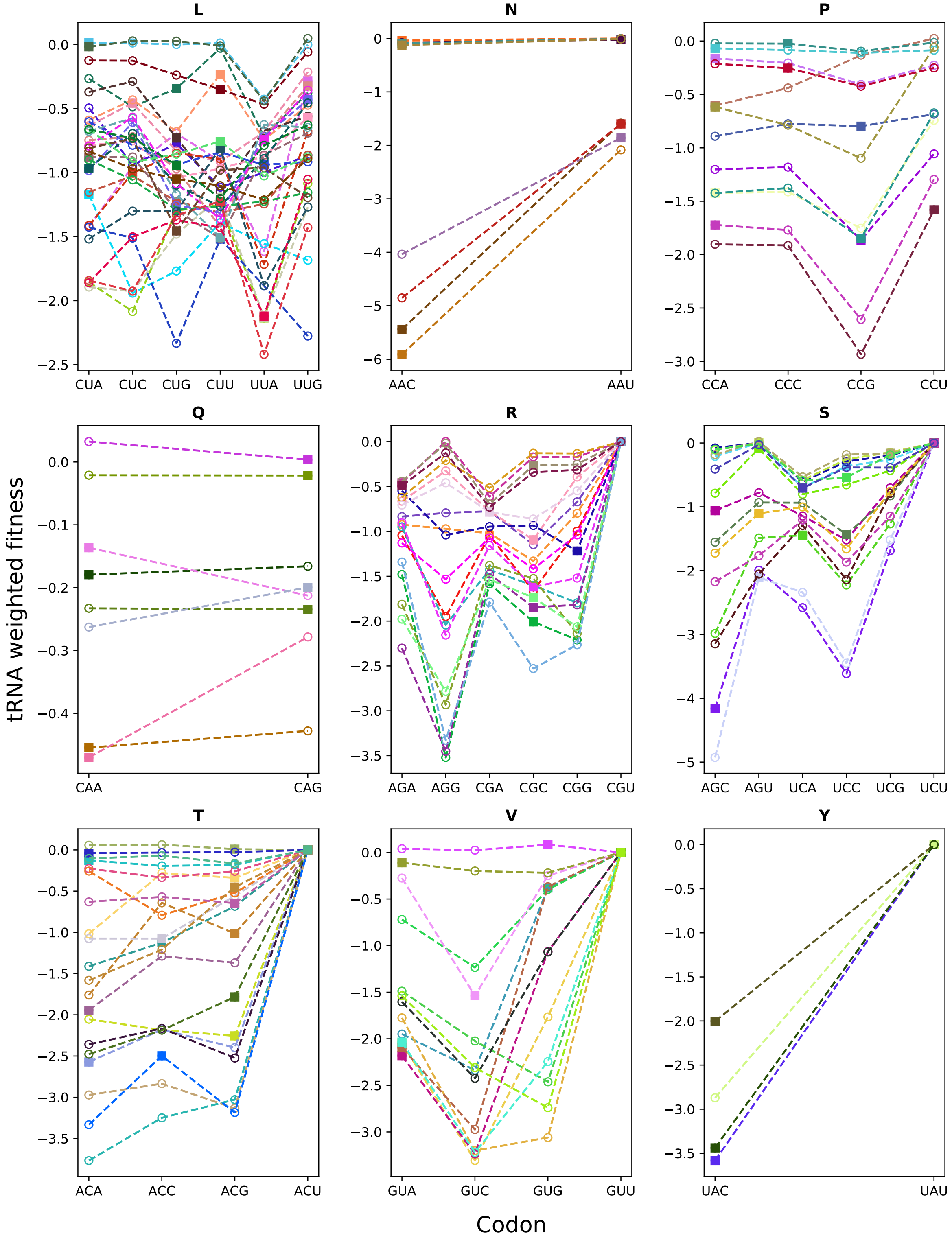

Figure 9.2: See caption of Figure 9.1.

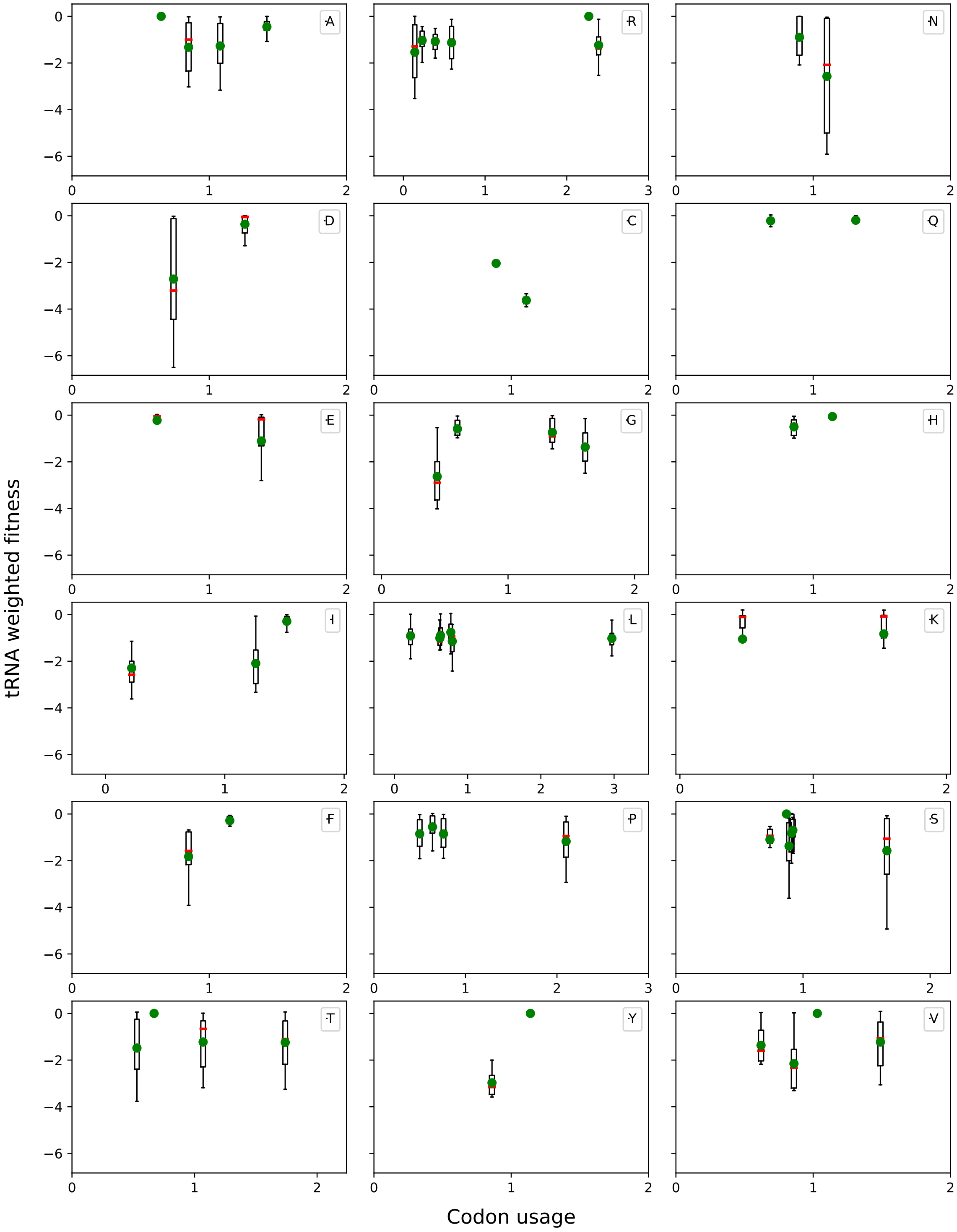

Figure 9.3: Distributions of average of the tRNA weighted fitness plotted with respect to codon usage for the possible codons of each amino acid for β-lactamase. Codon usage is defined as the ratio of number of observed occurrence to that of the expected if all codons for an amino acid were used equally. If fitness effects influence codon usage significantly an increase in fitness with the increase in codon usage is expected.

two criteria we used for choosing the optimal codon were found to be complementary, and of significant value together, we could not identify which criterion suits which amino acid. Whether a combination of both these criteria into an artificial intelligence model, or adding additional details such as the nucleotide substitution probability or the recognition of mRNA by wobbling with the tRNA will improve the predictability of the site-specific codon bias will be explored in the future.

9.3 Conclusion

We defined a new site-specific criterion for codon usage bias. This criterion uses deep mutational scan data from non-synonymous mutations, and rank orders the synonymous codons at each amino acid position in the protein using the potential loss of fitness upon a single nucleotide substitution. Interestingly the codons identified by this minimization of loss of fitness effect were complementary to those predicted by tRNA availability. Additional factors such as wobble-pairing, chance of substitution, effect of multiple substitutions have to be considered in a future work to see if the predictive capacity for the choice of the codon can be improved.

Bibliography

[1] R. Grantham, C. Gautier, M. Gouy, R. Mercier, and A. Pave, "Codon catalog usage and the genome hypothesis," *Nucleic acids research*, vol. 8, no. 1, pp. 197–197, 1980.

[2] P. M. Sharp and W.-H. Li, "An evolutionary perspective on synonymous codon usage in unicellular organisms," *Journal of molecular evolution*, vol. 24, no. 1-2, pp. 28–38, 1986.

[3] T. Ikemura, "Codon usage and trna content in unicellular and multicellular organisms.," *Molecular biology and evolution*, vol. 2, no. 1, pp. 13–34, 1985.

[4] H. Akashi, "Inferring weak selection from patterns of polymorphism and divergence at" silent" sites in drosophila dna.," *Genetics*, vol. 139, no. 2, pp. 1067–1076, 1995.

[5] H. Akashi, R. M. Kliman, and A. Eyre-Walker, "Mutation pressure, natural selection, and the evolution of base composition in drosophila," in *Mutation and Evolution*, pp. 49–60, Springer, 1998.

[6] M. Bulmer, "The selection-mutation-drift theory of synonymous codon usage.," *Genetics*, vol. 129, no. 3, pp. 897–907, 1991.

[7] S. L. Chen, W. Lee, A. K. Hottes, L. Shapiro, and H. H. McAdams, "Codon usage

between genomes is constrained by genome-wide mutational processes," *Proceedings of the National Academy of Sciences*, vol. 101, no. 10, pp. 3480–3485, 2004.

[8] S. Kanaya, Y. Yamada, M. Kinouchi, Y. Kudo, and T. Ikemura, "Codon usage and trna genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with cg-dinucleotide usage as assessed by multivariate analysis," *Journal of molecular evolution*, vol. 53, no. 4-5, pp. 290–298, 2001.

[9] M. Gouy and C. Gautier, "Codon usage in bacteria: correlation with gene expressivity," *Nucleic acids research*, vol. 10, no. 22, pp. 7055–7074, 1982.

[10] S. Kanaya, Y. Yamada, Y. Kudo, and T. Ikemura, "Studies of codon usage and trna genes of 18 unicellular organisms and quantification of bacillus subtilis trnas: gene expression level and species-specific diversity of codon usage based on multivariate analysis," *Gene*, vol. 238, no. 1, pp. 143–155, 1999.

[11] G. Cannarozzi, N. N. Schraudolph, M. Faty, P. von Rohr, M. T. Friberg, A. C. Roth, P. Gonnet, G. Gonnet, and Y. Barral, "A role for codon order in translation dynamics," *Cell*, vol. 141, no. 2, pp. 355–367, 2010.

[12] G. A. Gutman and G. W. Hatfield, "Nonrandom utilization of codon pairs in escherichia coli," *Proceedings of the National Academy of Sciences*, vol. 86, no. 10, pp. 3699–3703, 1989.

[13] J. R. Buchan, L. S. Aucott, and I. Stansfield, "trna properties help shape codon pair preferences in open reading frames," *Nucleic acids research*, vol. 34, no. 3, pp. 1015–1027, 2006.

[14] X. Xia, "Maximizing transcription efficiency causes codon usage bias," *Genetics*, vol. 144, no. 3, pp. 1309–1320, 1996.

[15] A. Lazrak, L. Fu, V. Bali, R. Bartoszewski, A. Rab, V. Havasi, S. Keiles, J. Kappes, R. Kumar, E. Lefkowitz, *et al.*, "The silent codon change i507-atc? att contributes to the severity of the δf508 cftr channel dysfunction," *The FASEB Journal*, vol. 27, no. 11, pp. 4630–4645, 2013.

[16] V. Presnyak, N. Alhusaini, Y.-H. Chen, S. Martin, N. Morris, N. Kline, S. Olson, D. Weinberg, K. E. Baker, B. R. Graveley, *et al.*, "Codon optimality is a major determinant of mrna stability," *Cell*, vol. 160, no. 6, pp. 1111–1124, 2015.

[17] S. Pechmann and J. Frydman, “Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding,” *Nature structural & molecular biology*, vol. 20, no. 2, p. 237, 2013.

[18] B. R. Morton, “Selection at the amino acid level can influence synonymous codon usage: implications for the study of codon adaptation in plastid genes,” *Genetics*, vol. 159, no. 1, pp. 347–358, 2001.

[19] P. Błażej, D. Mackiewicz, M. Wnętrzak, and P. Mackiewicz, “The impact of selection at the amino acid level on the usage of synonymous codons,” *G3: Genes, Genomes, Genetics*, vol. 7, no. 3, pp. 967–981, 2017.

[20] C. R. Woese, “On the evolution of the genetic code.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 54, no. 6, p. 1546, 1965.

[21] C. J. Epstein, “Role of the amino-acid ?code?and of selection for conformation in the evolution of proteins,” *Nature*, vol. 210, no. 5031, pp. 25–28, 1966.

[22] T. Sonneborn, “Degeneracy of the genetic code: extent, nature, and genetic implications,” in *Evolving genes and proteins*, pp. 377–397, Elsevier, 1965.

[23] M. Archetti, “Codon usage bias and mutation constraints reduce the level of errorminimization of the genetic code,” *Journal of Molecular Evolution*, vol. 59, no. 2, pp. 258–266, 2004.

[24] E. Firnberg, J. W. Labonte, J. J. Gray, and M. Ostermeier, “A comprehensive, high-resolution map of a gene?s fitness landscape,” *Molecular biology and evolution*, vol. 31, no. 6, pp. 1581–1592, 2014.

[25] M. A. Stiffler, D. R. Hekstra, and R. Ranganathan, “Evolvability as a function of purifying selection in tem-1 β-lactamase,” *Cell*, vol. 160, no. 5, pp. 882–892, 2015.

[26] A. Melnikov, P. Rogov, L. Wang, A. Gnirke, and T. S. Mikkelsen, “Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes,” *Nucleic acids research*, vol. 42, no. 14, pp. e112–e112, 2014.

[27] P. M. Sharp, L. R. Emery, and K. Zeng, “Forces that influence the evolution of codon bias,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 365, no. 1544, pp. 1203–1212, 2010.

[28] H. Akashi, “Synonymous codon usage in drosophila melanogaster: natural selection and translational accuracy.,” *Genetics*, vol. 136, no. 3, pp. 927–935, 1994.

[29] N. Stoletzki and A. Eyre-Walker, "Synonymous codon usage in escherichia coli: selection for translational accuracy," *Molecular biology and evolution*, vol. 24, no. 2, pp. 374–381, 2007.

[30] M. Orešič and D. Shalloway, "Specific correlations between relative synonymous codon usage and protein secondary structure," *Journal of molecular biology*, vol. 281, no. 1, pp. 31–48, 1998.

Chapter 10

Conclusions and Future Outlook

The advent of next generation sequencing technologies has made DNA sequencing to be cheaper, faster and more accurate. This development has provided access to very valuable genetic information, as exhaustive as the whole genome sequencing to identifying the single nucleotide polymorphisms implicated in diseases. Specifically in this thesis we investigate several conceptual questions on predicting and interpreting the effects of mutations. We used data such as large sequence alignments of rapidly mutating viruses and libraries of mutational effects to ask questions whose relevance will increase with increasing availability of data.

Viral complexity. We explored if the complexity of a viral genome can be defined using the mutational patterns obtained from large sequence alignments. Within the data available, the differences in the density of amino acid covariance network were seen to be correlated with the biological complexity quantified as the number of mortalities. But this surprising pattern raises more questions. The number of complete viral genome data available, varied between 1000 to 8000, which is an order of magnitude higher than about a decade ago, which can be expected to increase much more in the coming years. When more data on several other viruses will be available, it will be interesting to see whether the correlation we observed between co-evolutionary networks and biological complexity persists. Would it be possible to design drugs based on the clusters of co-varying amino acids? In case of pandemics, would it be possible to obtain enough sequences, sort them by the time of collection to see if there are any quantitative trends that show a transition from a lethal to a manageable infection?

Directed effects in amino acid coevolution. Pairs and clusters of symmetric amino acid interactions have been extremely useful in identifying hotspot residues in proteins. But most relations, including co-evolutionary relations are asymmetric, and hence we introduced a way of defining and quantifying this asymmetry using amino acid impact factor. We could identify a few critical relationships among amino acids, and their functional roles. These asymmetric relations identified from mutations raise further questions on whether they could be helpful in identifying the directionality of dynamical or functional relationships among the amino acids in wild type proteins, and if they are useful in identifying the pathways of allosteric communication in the proteins.

Sequence, structure, or dynamics. Directed or undirected, how much of the amino acid correlation information obtained independently through sequence, structure and dynamics based approaches is common? Conceptually they should all contain the same information. However, possibly because none of the methods are complete by themselves, the important amino acids that are selected using any of the methods turn out to be different from the ones selected using other approaches. In which case, how does one combine these three sets of information? It remains to be seen whether using the three sets of information combined using simple rules or artificial intelligence can be helpful in predicting the important amino acids in the protein which can or can not tolerate mutations.

Deep2Full. In the context of large scale mutagenesis data we asked if computational models can be built to reduce the number of variants characterised. What we performed using data on single mutations, where the predictions were tested against the measured data serves as a validation at this stage. As the emphasis shifts towards double, triple mutants, the hybrid models which combine partial experimental data with artificial intelligence models will become very valuable. While working with multiple mutants, the additive or compensatory effects are hard to predict. It will be interesting to explore whether some of the parameters we computed from the symmetric and asymmetric co-evolutionary patterns capture these compensatory effects and if they do, how we can build a reliable genotype to phenotype model for the mutational effects in proteins.

Interpreting mutational effects. In working with computational or theoretical models, transparency of how prediction is made is as important as the prediction. Especially while working with artificial intelligence models, this transparency can serve to interpret the physical basis of the predictions or at least clarify how the predictions are made, and possibly help learn more about the system. To achieve transparency in mutational effects predictions, we worked with different approaches, from developing simple rules of thumb, to a decision tree that can be tracked with ease to using methods of interpretable artificial intelligence. These approaches attempt to take the computational predictions and large sets of experimental data beyond the libraries they are, to initiate a dialogue with the biochemists. Biochemists have intuitions about relation between mutational effects and structural and evolutionary information about the protein. How does this new class of experiments which can probe tens of thousands of mutations or artificial intelligence approaches which can predict most or all of these mutations add to what the biochemists already know? With the new experiments, new analyses, are we strengthening the intuitions or are there any new learnings? These questions become interesting as more experiments with large scale mutational data, with reliable range of phenotypic effects become available.

Codon usage bias. Mutational effects libraries can be useful in understanding other cellular phenomena as well such has been attempted in predicting the choice of synonymous codon. Evolutionary models considering the selection at the amino acid level were based on general substitution matrices and can comment only about the codon usage at the complete genome level. But with the deep mutational scan data set we can explore the codon preferences at each amino acid position. A computational model to predict the same considering other selection pressures also as inputs can be developed when more data become available.

The world of proteins is fascinating, and working with data from large scale experiments and advanced algorithms for effects predictions opens up the possibility to ask interesting questions that were not possible earlier. While the sequence or mutational effects data is orders of magnitude larger than what was available a decade ago, it is clear that the fields exploring the functional roles of proteins using next generation sequencing are still nascent. We hope the questions we asked in the thesis will add to the repertoire of the next generation studies on understanding proteins.

Appendix A

Statistical Characteristics of Amino Acid Covariance as Possible Descriptors of Viral Genomic Complexity

Protein	GAG	POL	VIF	VPR	TAT	REV	VPU	ENV	NEF
(No. of amino acids)	(500)	(1003)	(192)	(96)	(100)	(116)	(82)	(856)	(205)
GAG	75	176	74	24	34	56	23	162	63
POL		180	154	43	56	110	56	317	133
VIF			37	23	23	46	17	146	54
VPR				1	9	11	5	36	10
TAT					9	34	8	95	27
REV						19	18	112	44
VPU							8	74	24
ENV								226	140
NEF									53

Table A.1: Table showing the number of inter-protein and intra-protein amino acid covariance relations from HIV data, with $C^{th} = 0.7$.

Protein (No. of amino acids)	NP	PB2	HA	M1	M2	NA	NS1	NEP	PA	PB1-F2	PB1
	(498)	(759)	(566)	(252)	(97)	(469)	(230)	(121)	(716)	(90)	(757)
NP	30	42	51	118	65	7	168	16	91	94	83
PB2		40	29	129	58	2	163	35	115	69	104
HA			5641	164	26	1019	273	17	63	31	54
M1				130	144	16	353	42	206	118	174
M2					32	2	182	11	106	52	92
NA						6976	27	10	3	13	2
NS1							2694	1377	281	182	253
NEP								180	34	15	30
PA									148	93	163
PB1-F2										645	114
PB1											60

Table A.2: Table showing the number of inter-protein and intra-protein amino acid covariance couplings from avian influenza data, with $C^{th} = 0.7$.

Protein (No. of amino acids)	NP	PB2	HA	M1	M2	NA	NS1	NEP	PA	PB1-F2	PB1
	(498)	(759)	(565)	(252)	(97)	(470)	(230)	(121)	(716)	(57)	(757)
NP	572	918	3103	552	468	2259	1198	335	858	445	1031
PB2		371	2389	436	372	1803	932	275	710	475	850
HA			7218	1808	1509	8913	3876	702	2651	1087	2170
M1				153	259	1348	680	149	470	198	401
M2					94	1121	539	130	382	169	401
NA						3429	2916	533	1943	778	1625
NS1							688	354	968	462	931
NEP								48	246	143	273
PA									343	340	771
PB1-F2										566	330
PB1											429

Table A.3: Table showing the number of inter-protein and intra-protein amino acid covariance couplings from human influenza data, with $C^{th} = 0.7$.

Protein (No. of amino acids)	HBe (214)	HBc (185)	HBx (154)	LHBs (400)	MHBs (281)	SHBs (226)	Pol (845)	HBSP (113)
HBe	33	77	52	228	102	53	524	79
HBc		30	42	236	106	58	505	73
HBx			50	367	222	142	677	100
LHBs				850	986	660	3065	508
MHBs					253	383	1650	280
SHBs						114	1090	191
Pol							2698	917
HBSP								82

Table A.4: Number of inter-protein and intra-protein amino acid covariance couplings for hepatitis at $C^{th} = 0.7$.

Protein (No. of amino acids)	ancC (114)	M (75)	E (495)	NS1 (352)	NS2a (218)	NS2b (130)	NS3 (619)	NS4a (127)	k (23)	NS4b (249)	NS5 (899)
ancC	155	243	1206	740	726	358	1136	371	29	296	1737
M		83	897	527	495	256	861	250	31	222	1267
E			2177	2720	2527	1295	4133	1334	109	1092	6198
NS1				807	1607	826	2450	801	65	704	3638
NS2a					840	788	2273	837	56	656	3619
NS2b						178	1192	404	27	320	1833
NS3							1891	1205	109	1005	5683
NS4a								199	31	343	1845
k									1	23	155
NS4b										126	1525
NS5											4257

Table A.5: Table showing the number of inter-protein and intra-protein amino acid covariance couplings from dengue virus data, with a $C^{th} = 0.7$.

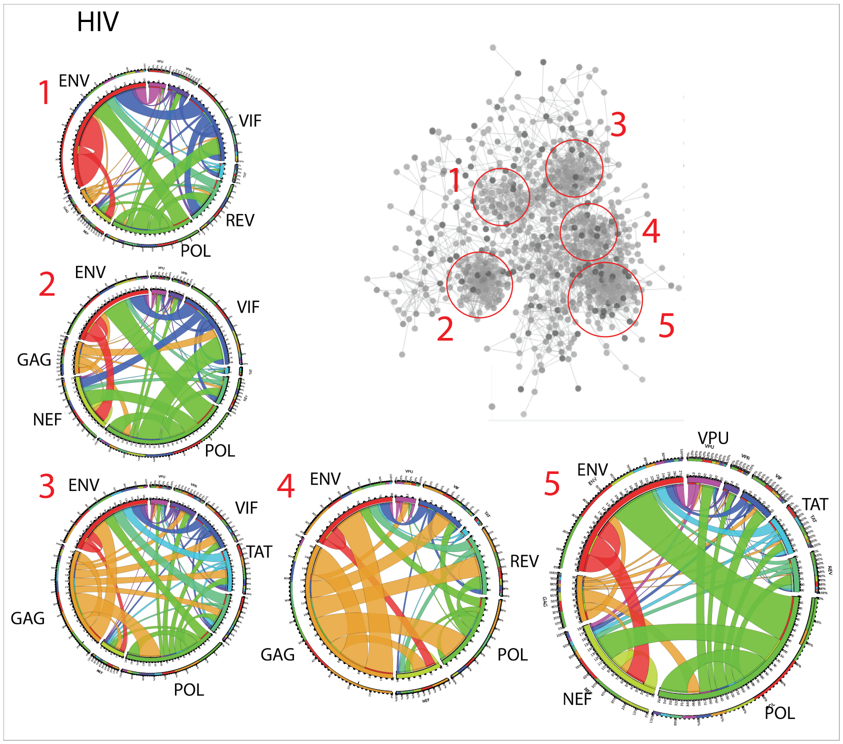

Figure A.1: Chord diagrams showing the strength of intra and inter protein interactions in each cluster of the covariance network of HIV. Size of the chord diagram is proportional to the number of amino acids in the cluster. Color indicates the protein. The network from Figure 2.1 of Chapter 2 is shown for reference. The proteins with most interactions are labeled in these chord diagrams.

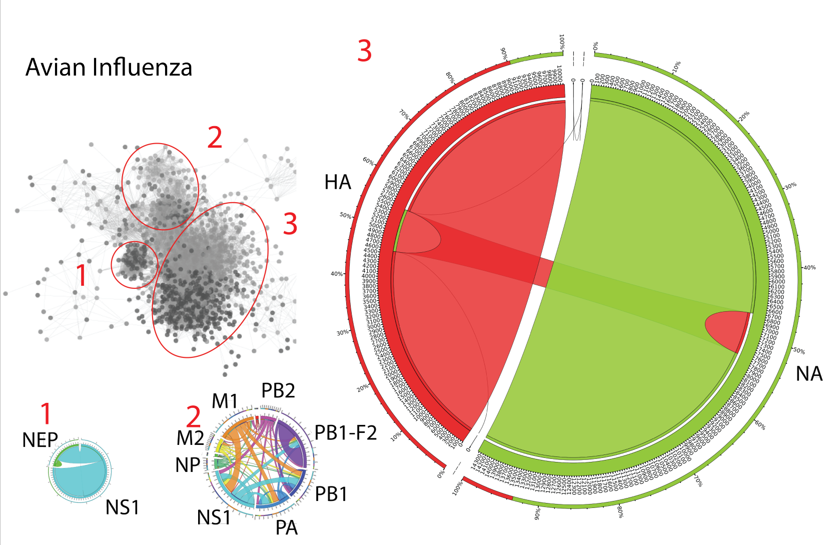

Figure A.2: Chord diagrams showing the strength of intra and inter protein interactions in each cluster of the covariance network of avian influenza. The size of the chord diagram is proportional to the number of amino acids in the cluster. Color indicates the protein. The network from Figure 2.1 of Chapter 2 is shown for reference. The proteins with most interactions are labeled in these chord diagrams.

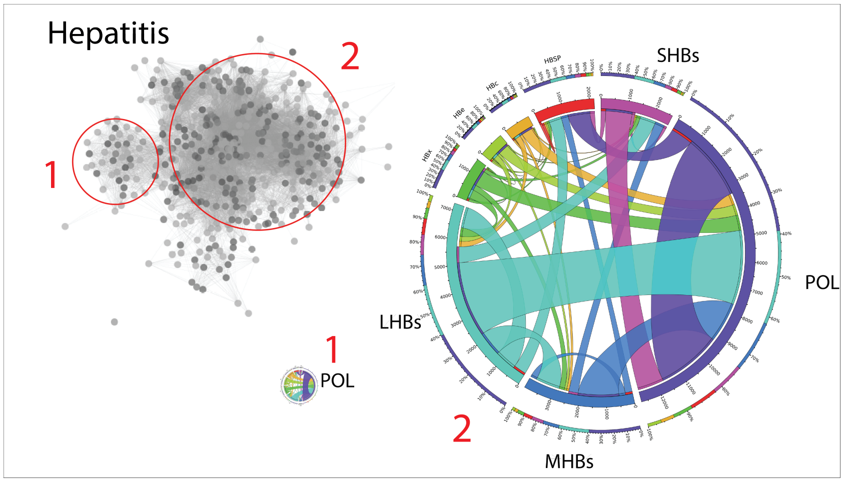

Figure A.3: Chord diagrams showing the strength of intra and inter protein interactions in each cluster of the covariance network of hepatitis. The size of the chord diagram is proportional to the number of amino acids in the cluster. Color indicates the protein. The network from Figure 2.1 of Chapter 2 is shown for reference. The proteins with most interactions are labeled in these chord diagrams.

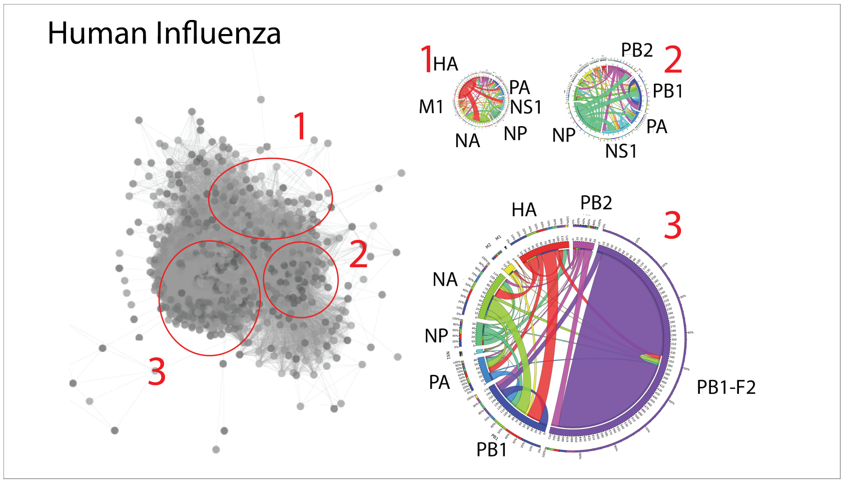

Figure A.4: Chord diagrams showing the strength of intra and inter protein interactions in each cluster of the covariance network of human influenza. The size of the chord diagram is proportional to the number of amino acids in the cluster. Color indicates the protein. The network from Figure 2.1 of Chapter 2 is shown for reference. The proteins with most interactions are labeled in these chord diagrams.

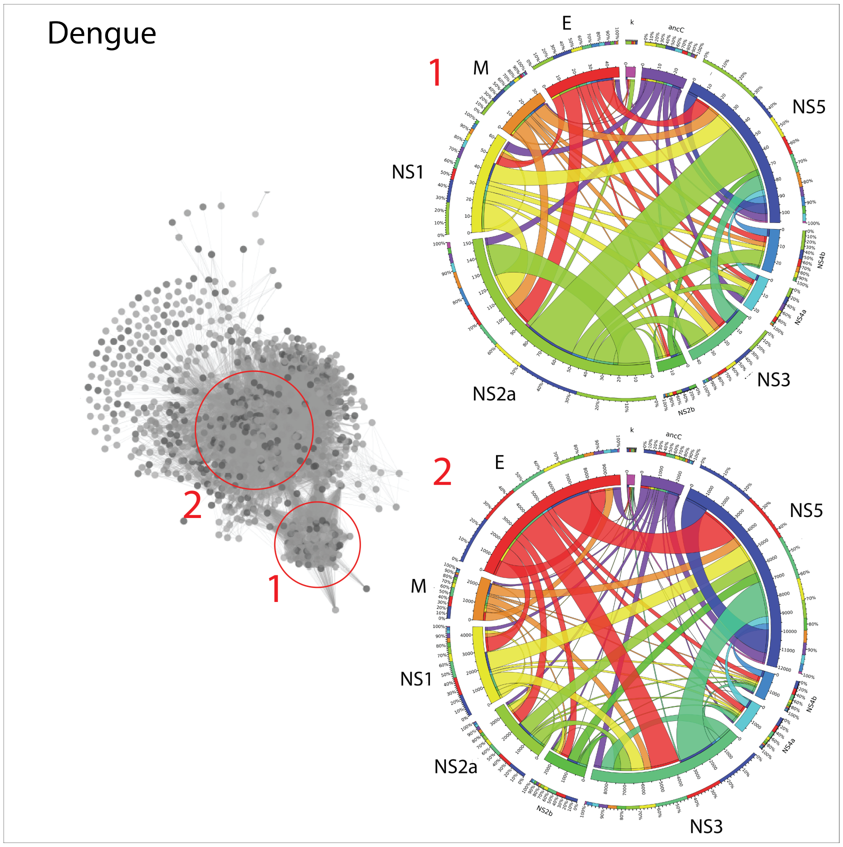

Figure A.5: Chord diagrams showing the strength of intra and inter protein interactions in each cluster of the covariance network of dengue. The size of the chord diagram is proportional to the number of amino acids in the cluster. Color indicates the protein. The network from Figure 2.1 of Chapter 2 is shown for reference. The proteins with most interactions are labeled in these chord diagrams.

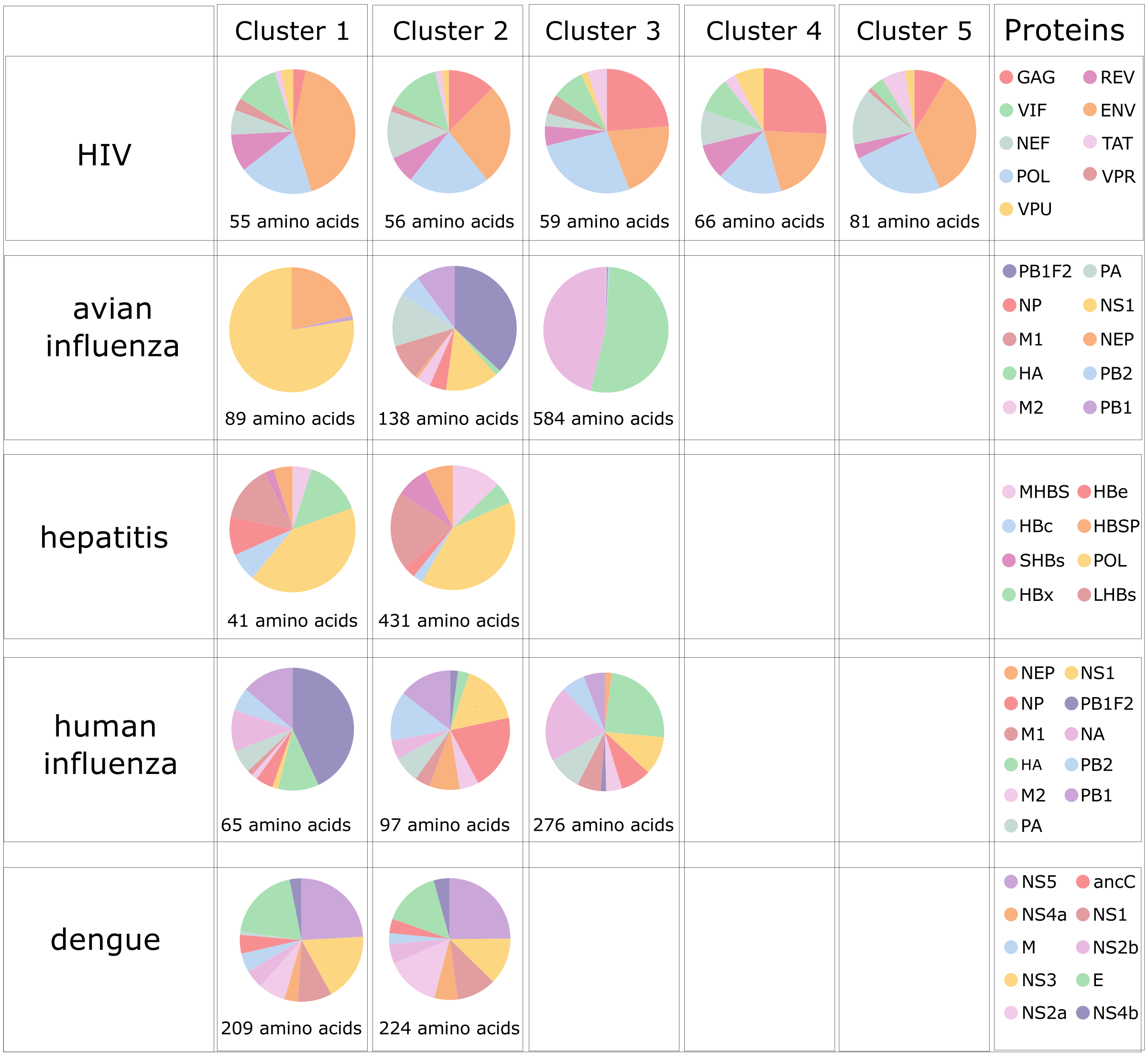

Figure A.6: Protein composition of clusters in the covariance network of all five viruses. The color scheme for proteins is indicated in the last column.

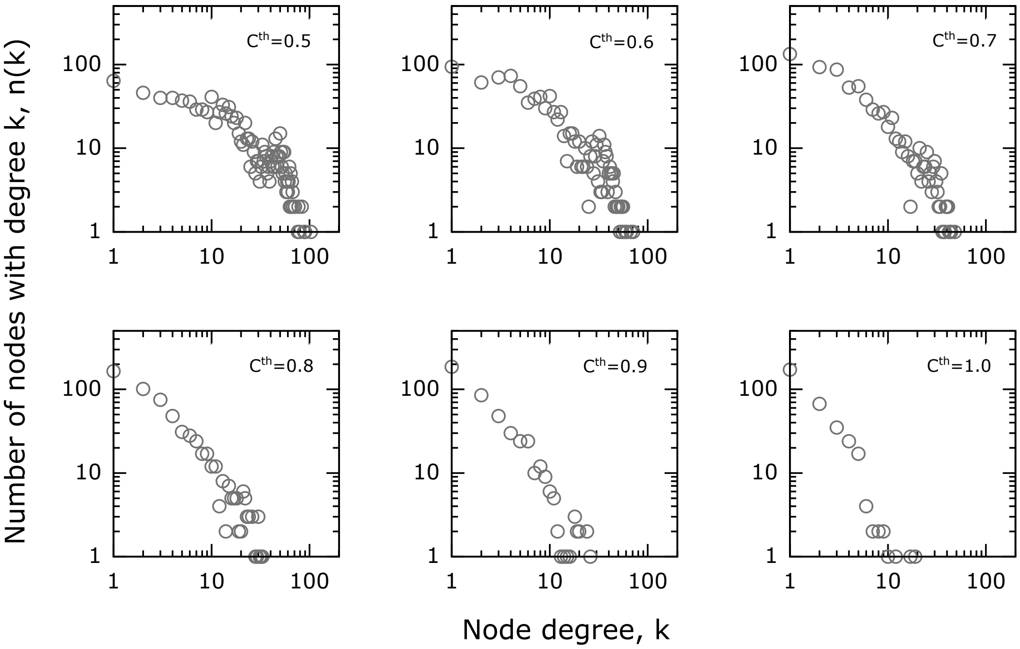

Figure A.7: Variation in the node degree distribution of HIV covariance network as the cutoff C^{th} is changed.

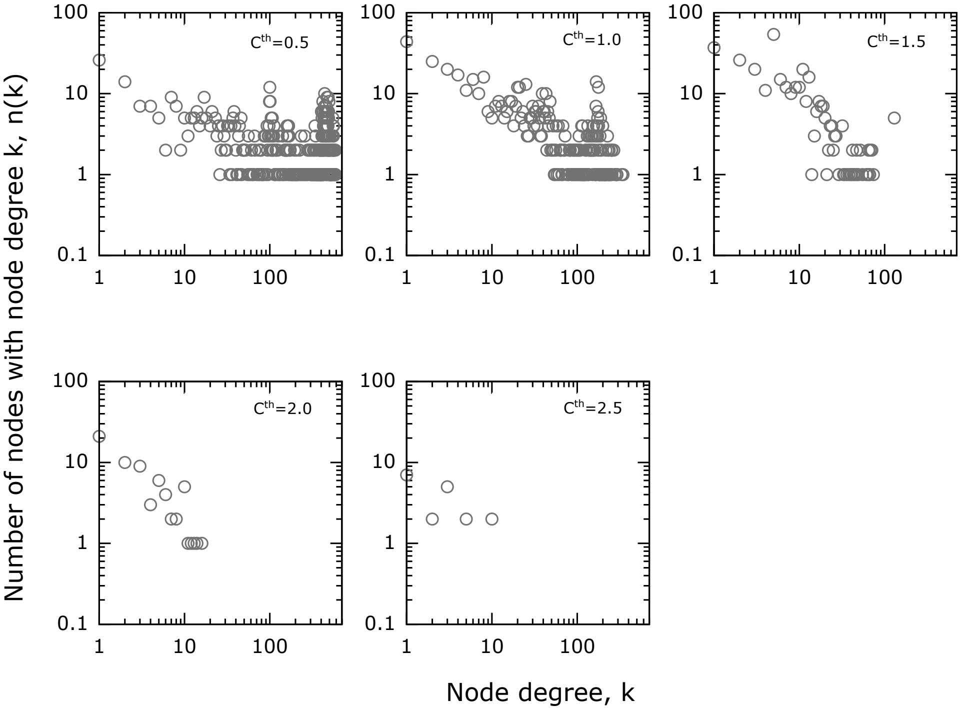

Figure A.8: Variation in the node degree distribution of human influenza covariance network as the cutoff C^{th} is changed.

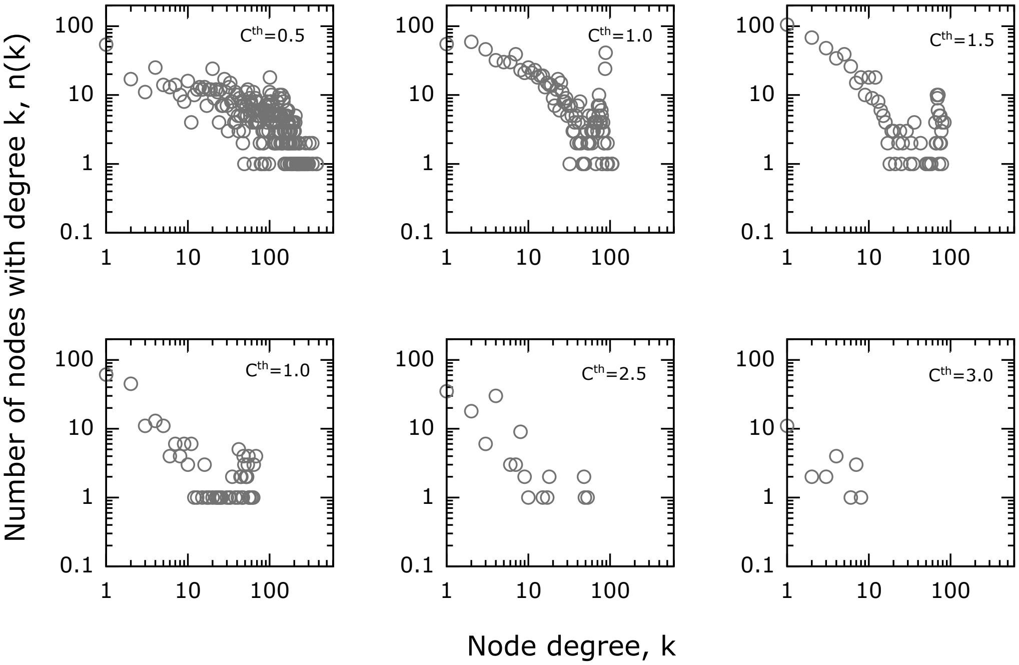

Figure A.9: Variation in the node degree distribution of avian influenza covariance network as the cutoff C^{th} is changed.

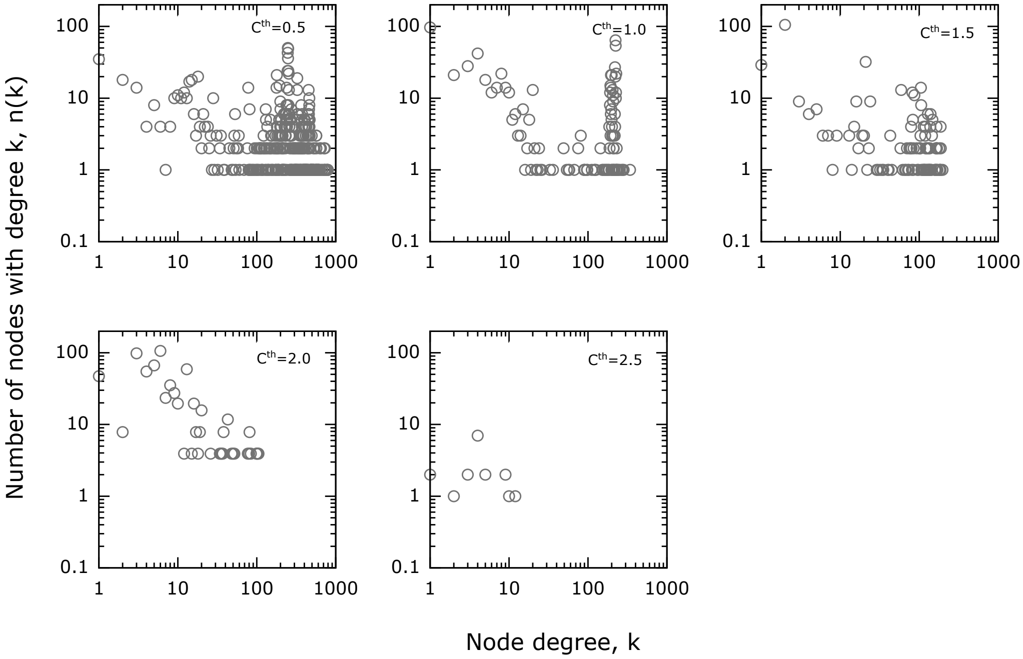

Figure A.10: Variation in the node degree distribution of dengue covariance network as the cutoff C^{th} is changed.

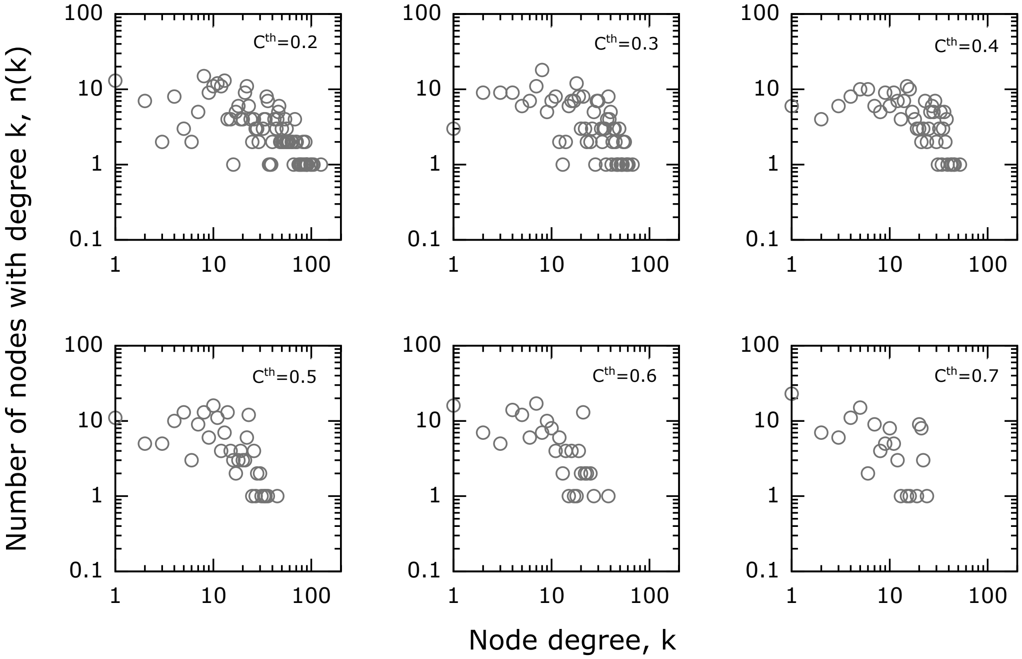

Figure A.11: Node degree distribution from the covariance network of dengue serotype 1. The analysis was performed on 1696 sequences, as a way of comparing the statistical behavior of one serotype with the combined serotype data.

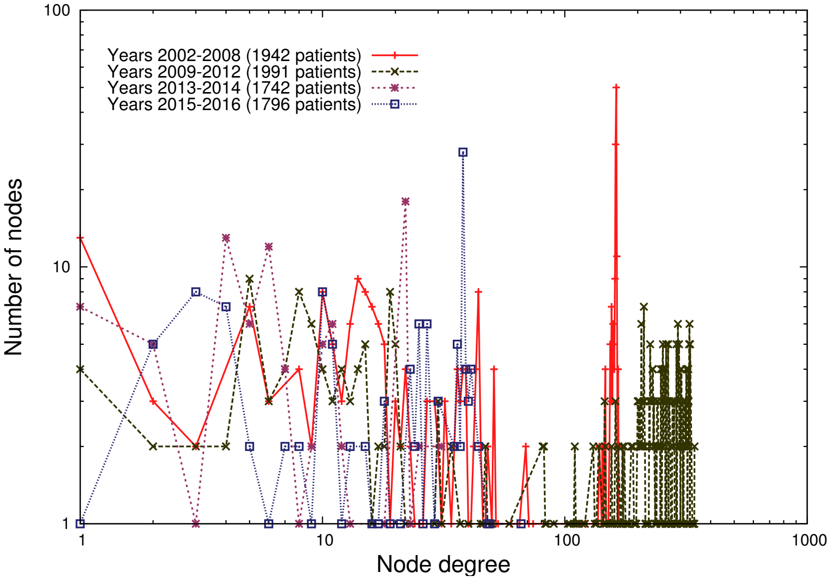

Figure A.12: Variation of the node degree distribution over years in the human influenza. Human influenza data was sorted according to the year of incidence and 4 groups of about 2000 patients each were made. No noticeable trend in the node degree distribution was observed in the data between 2002-2016.

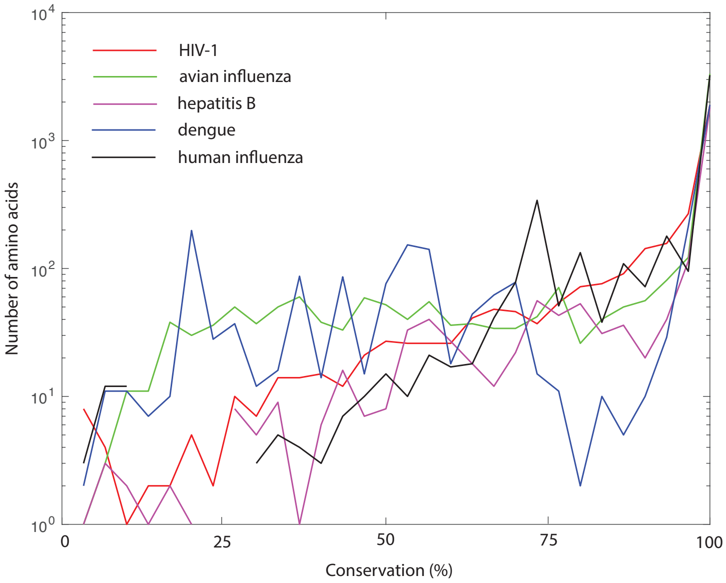

Figure A.13: Distribution of the conservation of amino acids in different viruses.

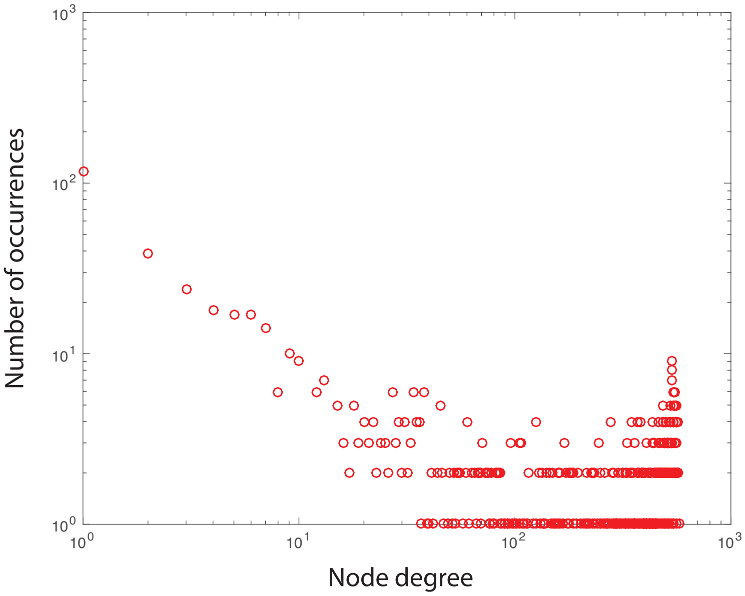

Figure A.14: Model network generated using the amino acid conservation distribution from HIV, and $\eta(\phi)$ with parameters $\phi_m = 0.05$ and $\sigma = 0.7$.

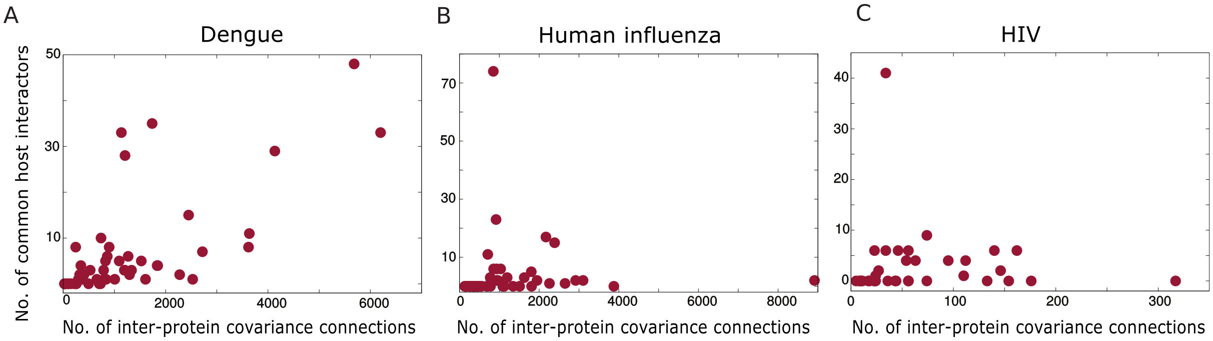

Figure A.15: A comparison of *pairwise* viral protein interaction strengths obtained from two different methods is shown. Number of inter-protein connections from our covariance analysis is compared with the number of host proteins commonly interacting with both the viral proteins, in (A) dengue (B) human influenza and (C) HIV.

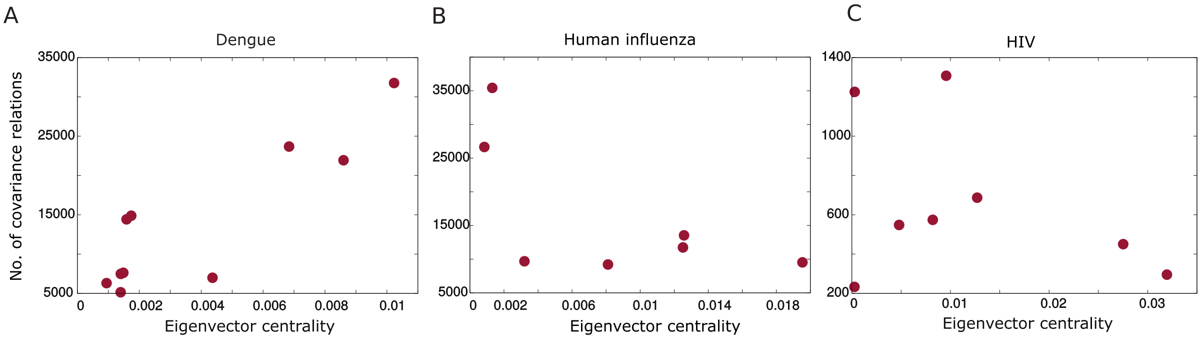

Figure A.16: A comparison of the relative importance of different viral proteins in our amino acid interaction network and in virus-host interactome is shown. Eigenvector centrality of the viral protein in the virus-host protein interactome is compared with the number of covariance relations the protein has for (A) dengue (B) human influenza and (C) HIV

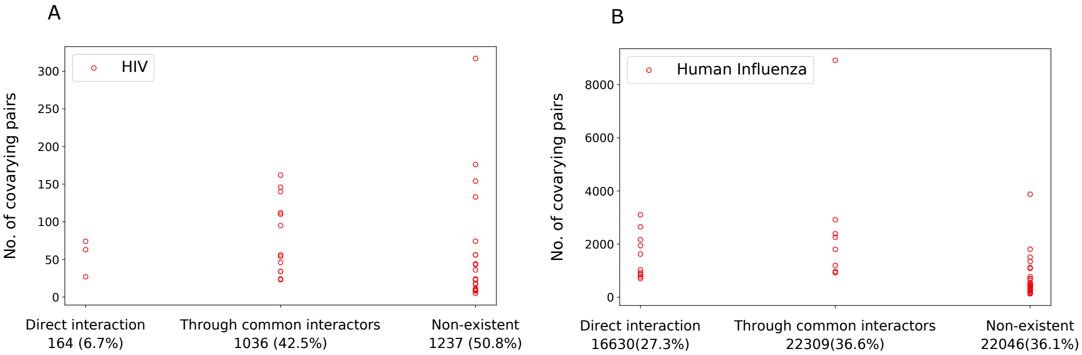

Figure A.17: Analysis of the viral inter-protein contacts, as direct, mediated by host proteins, or non-existent was performed where data was available. The results are shown for (A) HIV and (B) human influenza by plotting against the strength of inter-protein interactions from covariance analysis. No clear pattern was observed for the data available.

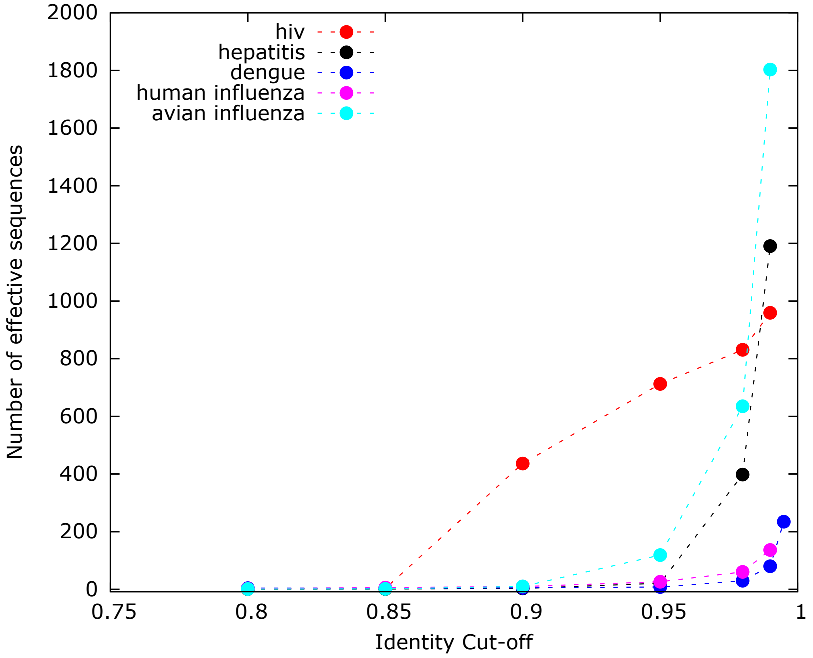

Figure A.18: The change in number of effective sequences with identity cut-off for the five viruses.

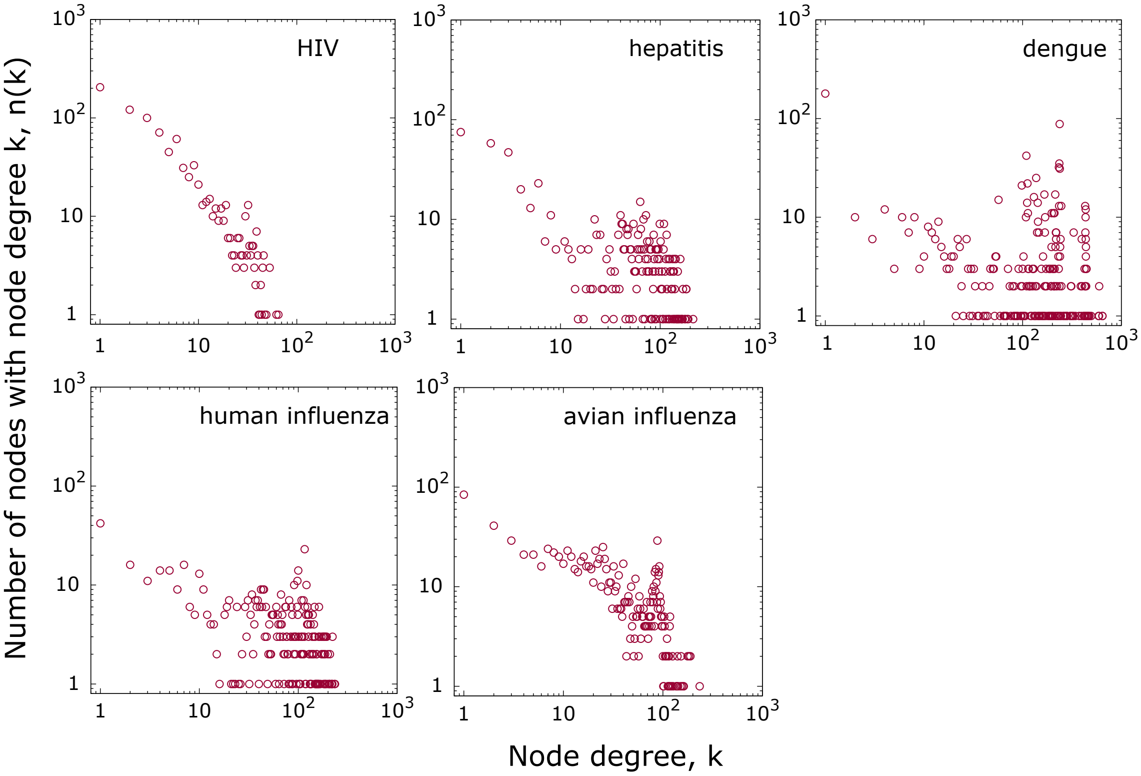

Figure A.19: The node degree distribution for HIV, hepatitis, dengue, human influenza and avian influenza for the covariance network generated using 200 randomly selected sequences from the complete data.

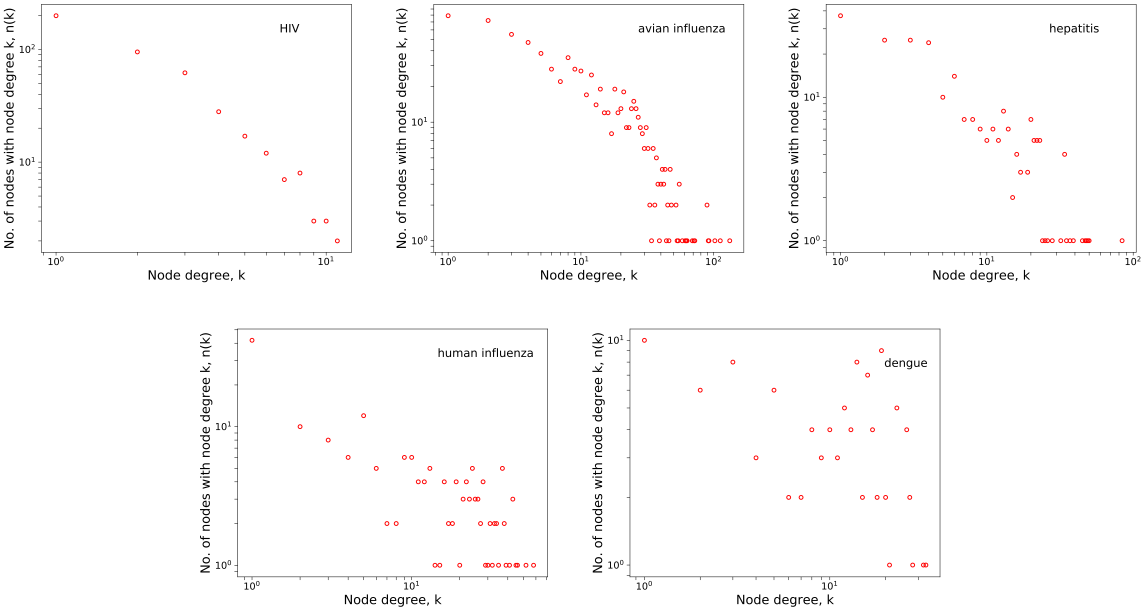

Figure A.20: Node degree distribution of networks generated from the covariance matrix after removing the contribution of top 5 eigen components.

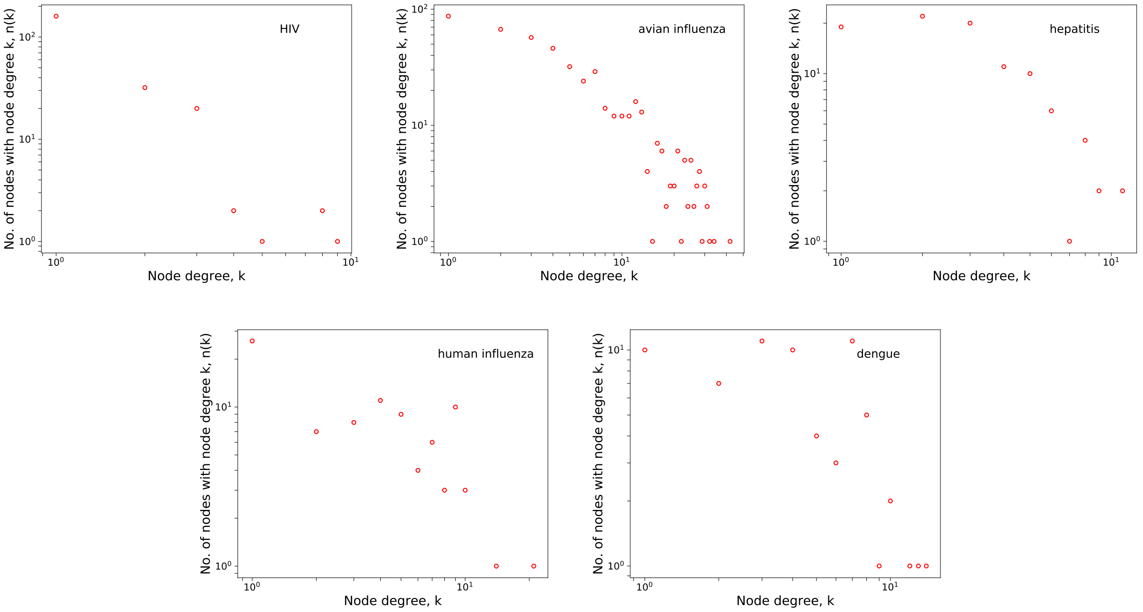

Figure A.21: Node degree distribution of networks generated from the covariance matrix after removing the contribution of top 10 eigen components. With the elimination of so many eigen components, the number of connections in the networks is reduced.

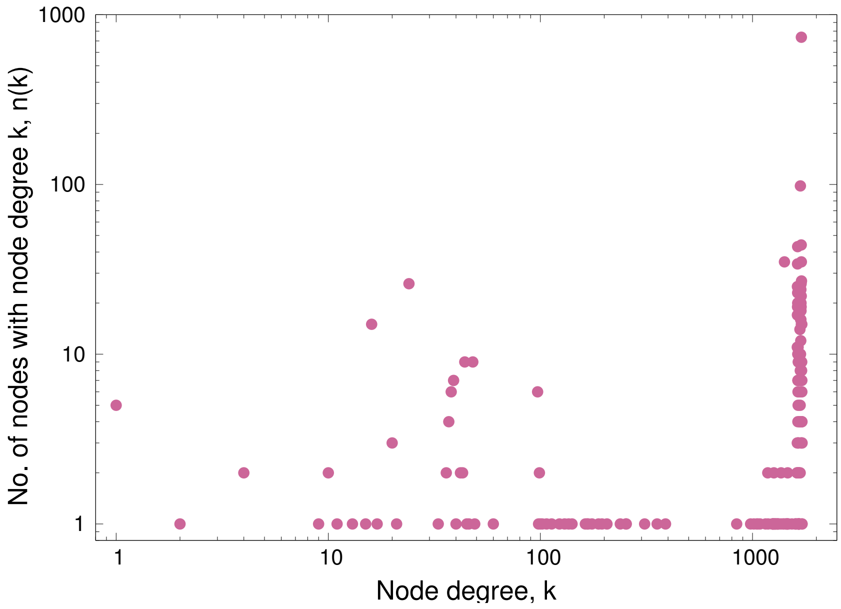

Figure A.22: Node degree distribution for the covariance network of dengue virus generated using MaxSubTree method ($C^{th} = 1.7$). The random nature of the distribution could be seen.

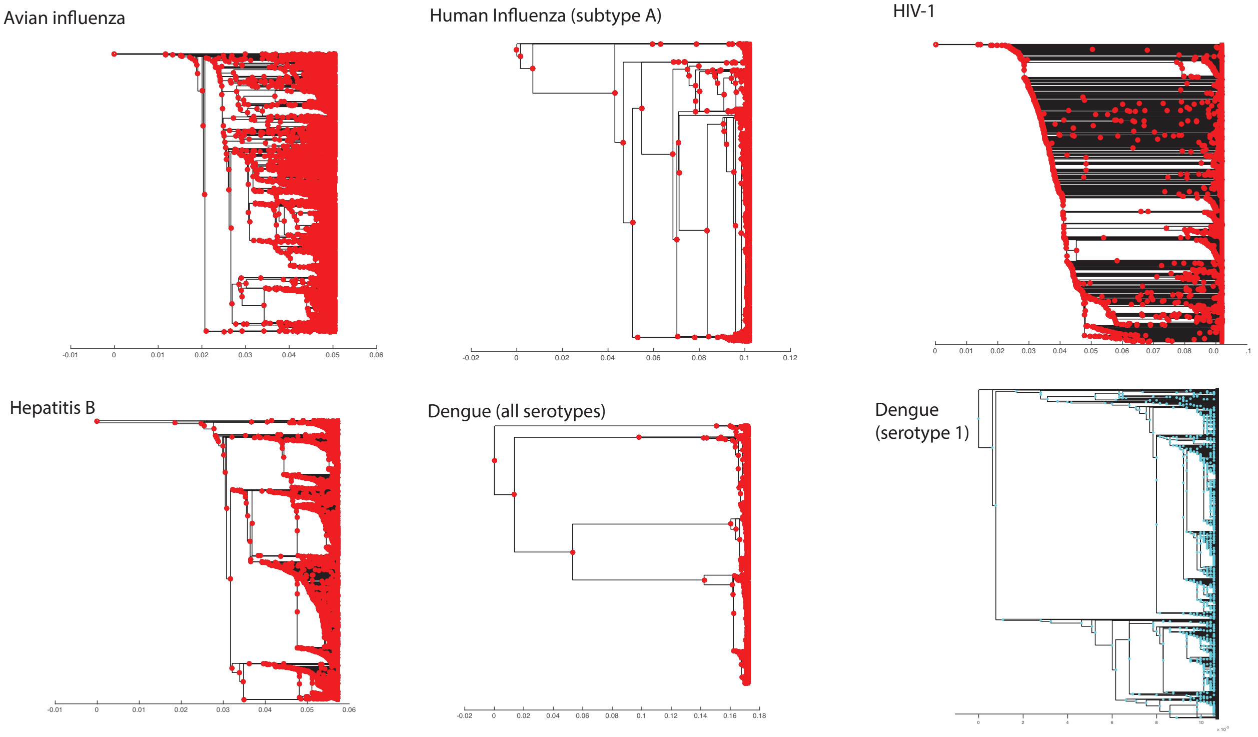

Figure A.23: Phylogenetic trees generated from the sequence data used for avian influenza, human influenza (subtype A) HIV-1 (subtype B), hepatitis B and dengue (all serotypes). The phylogenetic tree for dengue serotype 1 is shown as well.

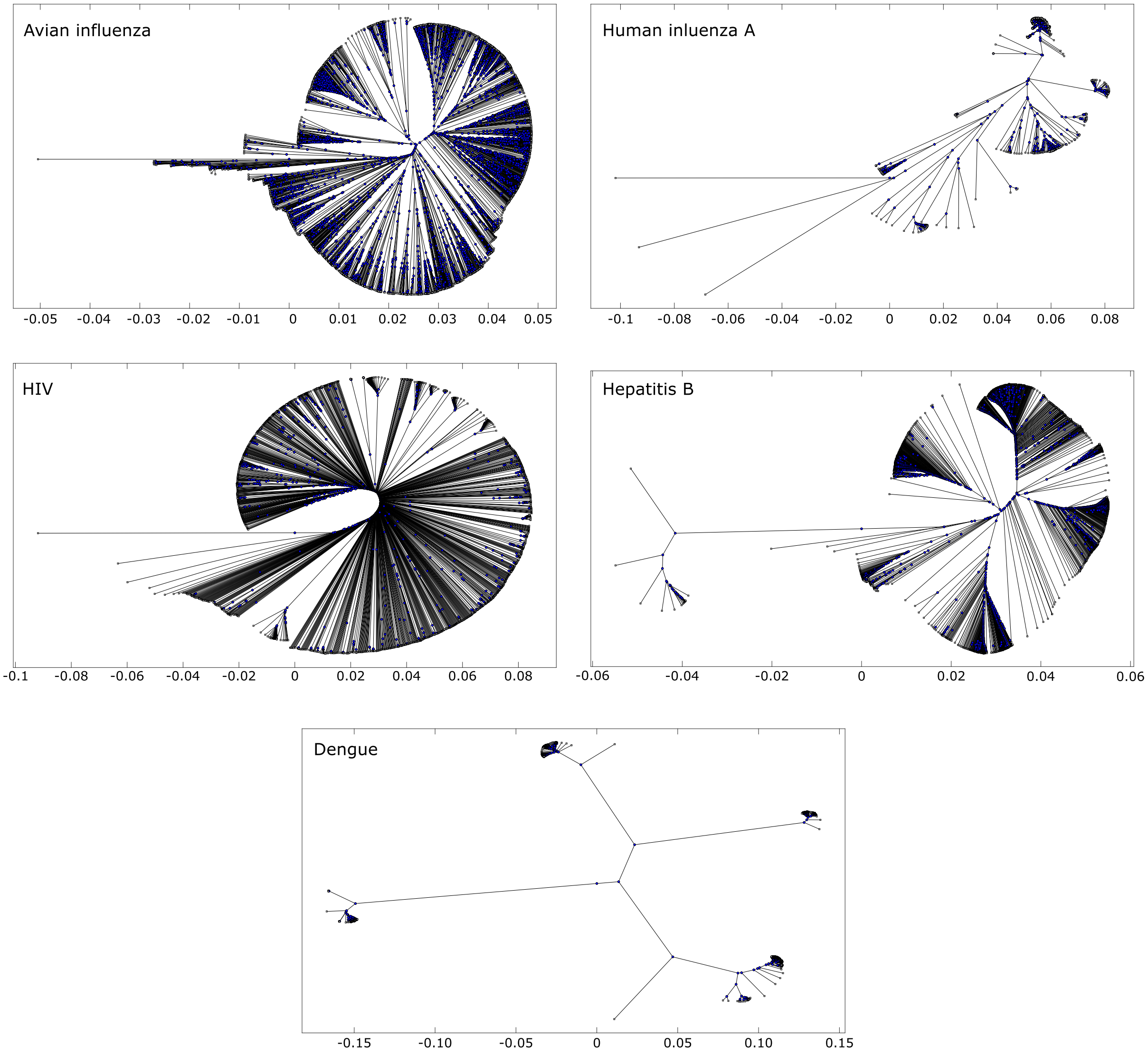

Figure A.24: Phylogenetic trees generated from the sequence data used for avian influenza, human influenza (subtype A), HIV-1 (subtype B) hepatitis B and dengue (all serotypes) in an unrooted representation.

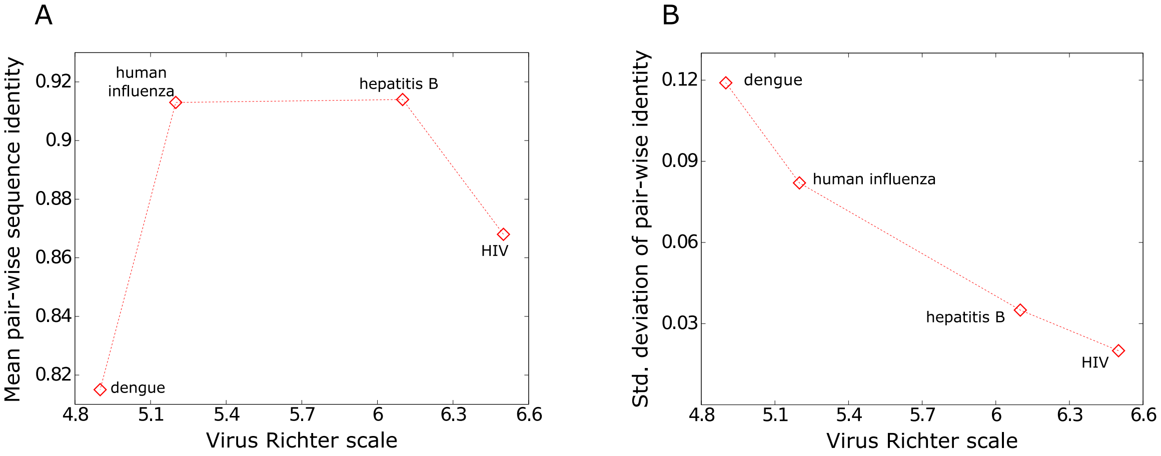

Figure A.25: Virus Richter scale versus (A) mean of the pair-wise sequence identities for the sequences in the alignment (B) standard deviation of the pair-wise sequence identities

Appendix B

Amino Acid Impact Factor

Residue Id	Impact factor Full data, $\gamma = 0.7$	Impact factor Full data, $\gamma = 0.8$	Impact factor Half data, $\gamma = 0.8$
19	2	-	-
29	1	1	1
32	1	1	1
34	2	-	-
40	1	1	1
42	1	-	-
57	1	1	1
58	1	1	1
100	1	-	-
102	2	1	1
122	1	-	-
136	1	1	1
140	3	-	-
142	2	-	-
168	1	-	-
182	2	-	1
183	2	-	-
184	2	-	-
189	1	1	1
191	1	-	-
194	3	-	-
196	8	1	-
201	1	1	1
211	1	-	-
216	2	-	-
226	1	1	1
228	2	-	-
237	1	-	-

Table B.1: Table showing the variation of impact factor with different data set sizes and with cut-offs 0.7 and 0.8 in the case of **serine protease.**

Protein	No. of residues included in the analysis	Reference sequence id	γ	Total no. of connections identified	No. of connections discarded because p-value > 0.01
Serine Protease	216	TRY2_RAT	0.7	48	0
DHFR	158	DYR_ECOLI	0.7	21	1
PGK	398	PGK1_HUMAN	0.8	361	0
HIV protease	99	K03455	0.8	44	16
HIV reverse transcriptase	440	K03455	0.8	340	182
GAG-POL polyprotein	1503	K03455	0.9	2486	1319

Table B.2: Table showing the number connections that are identified with a chosen γ as well as the number of connections with p-value > 0.01 that were discarded from the analysis. The Pfam alignment did not have all the residues in the reference PDB, so the analyses do not include all the sequence positions which are present in the pdb.

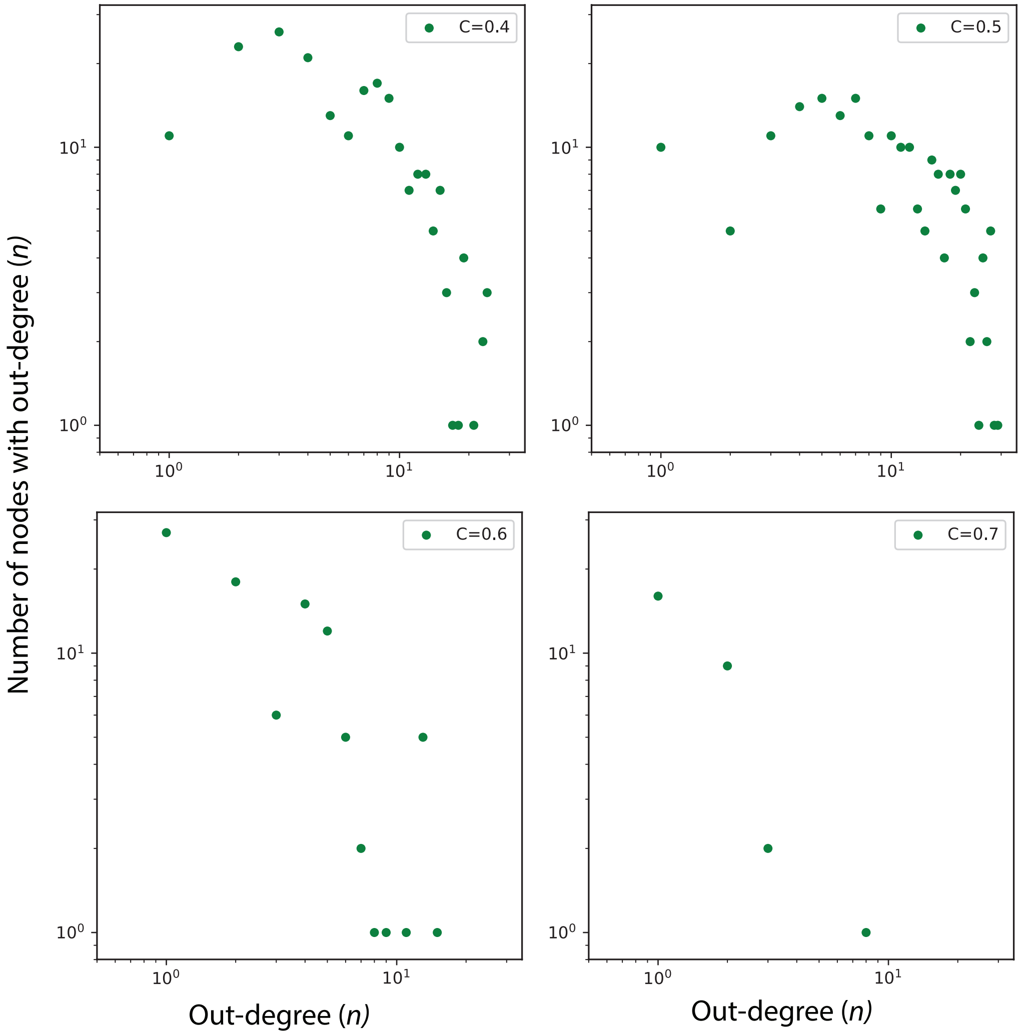

Figure B.1: The change in node-outdegree distribution for serine protease. The x-axis indicates the number of out-going connections from a node, and the y-axis shows how many such connections are present. The different subplots represent the same analysis performed with different choices of the cut-off. It can be seen that when the cut-off goes below 0.6, the network begins show a transition from an scale-free to random network.

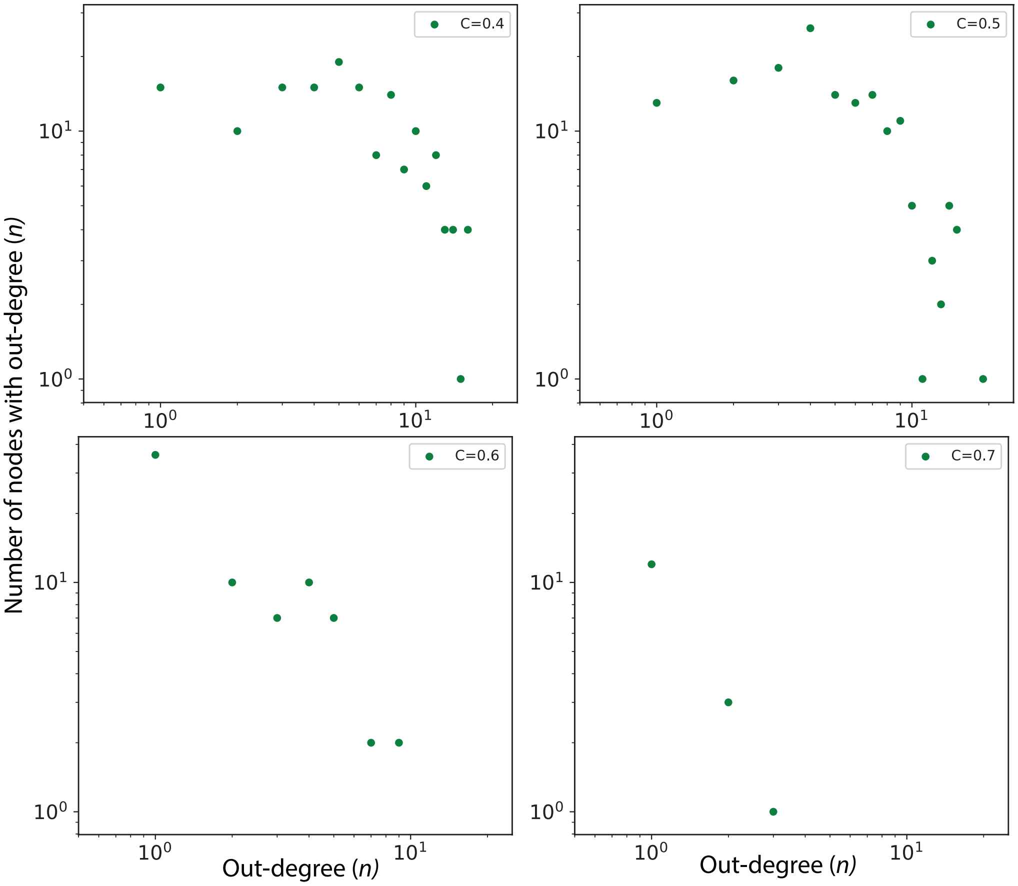

Figure B.2: The change in node-outdegree distribution for DHFR. The x-axis indicates the number of outgoing connections from a node, and the y-axis shows how many such connections are present. It can be seen that when the cut-off goes below 0.6, the network begins show a transition from an scale-free to random network.

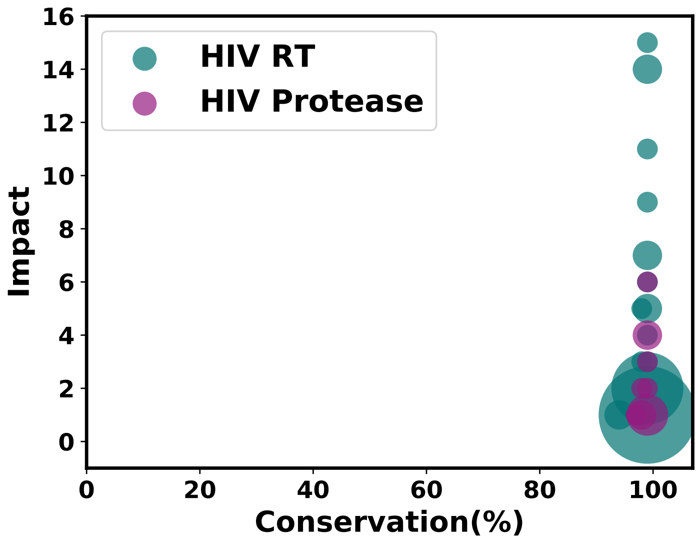

Figure B.3: Impact-Conservation analysis showing the residues with impact at $\gamma = 0.8$ on y-axis and conservation on x-axis for HIV-1 protease and reverse transcriptase.

Appendix C

Correlations from Structure, Sequence and Dynamics are Complementary Rather than Synonymous

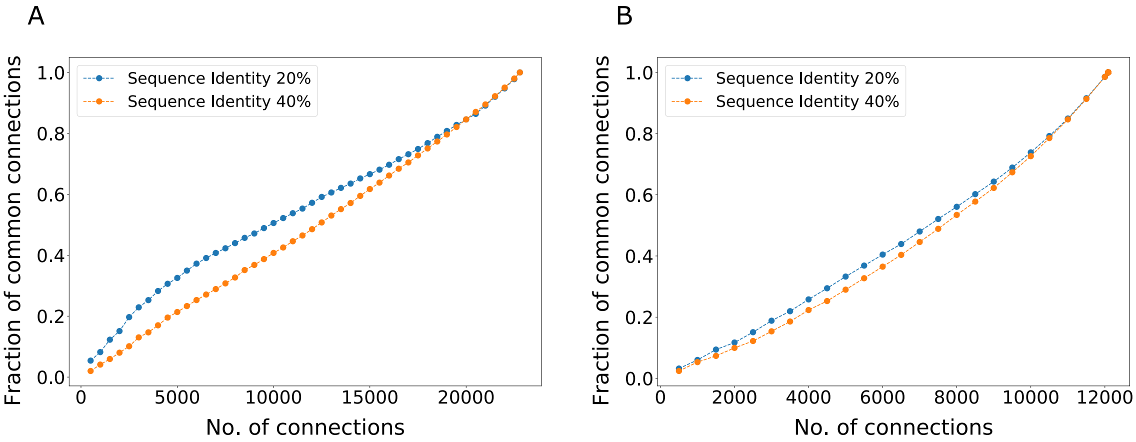

Figure C.1: Variation in the convergence when the sequence alignment consisted of sequences only with identity more 40%. For (A) Serine protease (B) DHFR.

Pair/Protein	Serine protease	DHFR
Sequence and Structure	-0.27	-0.15
Sequence and Dynamics	-0.22	-0.05
Structure and Dynamics	0.72	0.6

Table C.1: Pearson correlation between the sequence, structure and dynamics correlation matrices.

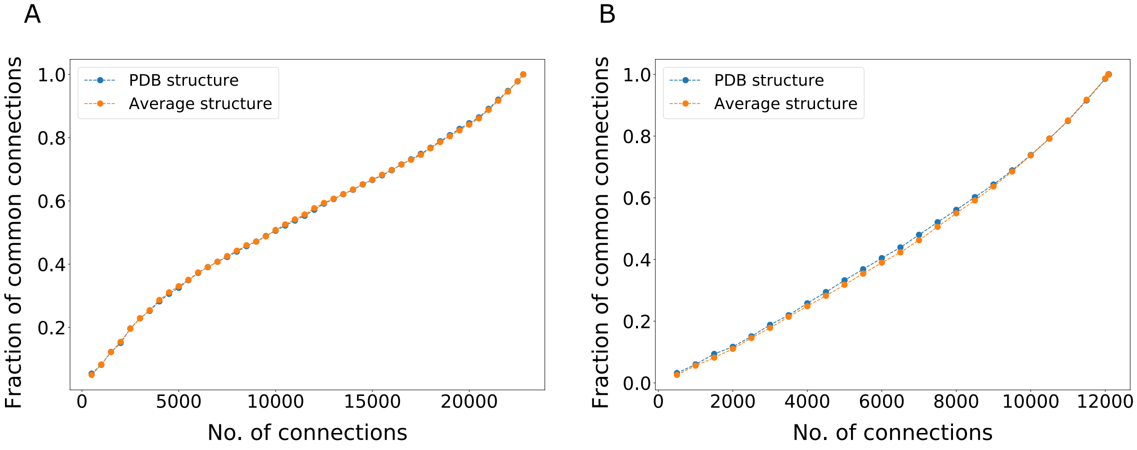

Figure C.2: Variation in the convergence on using the first 50 ns long MD trajectory comapred to the 100 ns. For (A) Serine protease (B) DHFR

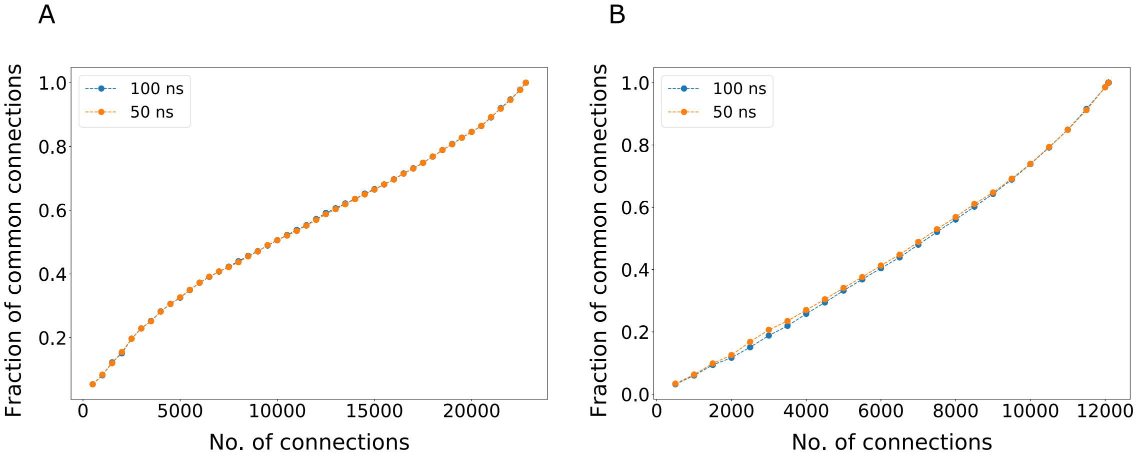

Figure C.3: Convergence of the connections identified from each of the sequence, structure and dynamics based approaches as the structural analysis was performed on the average structure obtained from the MD trajectory compared to the PDB structure for (A) Serine protease (B) DHFR

Residue No.	Residue	Analysis
55	A	seq,str,dyn
228	Y	seq,str,dyn
195	S	seq,str
197	G	seq,str
51	W	seq,dyn
31	V	str,dyn
42	C	str,dyn
44	G	str,dyn

Residue No.	Residue	Analysis
45	S	str,dyn
194	D	str,dyn
213	V	str,dyn
29	Y	seq
40	H	seq
57	H	seq
122	A	seq
124	P	seq
136	C	seq
141	W	seq
180	M	seq
189	D	seq
201	C	seq
215	W	seq
220	C	seq
225	P	seq
237	W	seq
238	I	seq
30	Q	str
43	G	str
102	D	str
117	R	str
142	G	str
196	G	str
198	P	str
199	V	str
211	G	str
212	I	str
46	L	dyn
52	V	dyn
53	V	dyn
54	S	dyn
105	L	dyn
106	I	dyn
200	V	dyn
209	L	dyn

Residue No.	Residue	Analysis
210	Q	dyn
227	V	dyn
229	T	dyn

Table C.2: Top 20 residues identified for serine protease using each of the approaches -sequence, structure and dynamics are given in the table. Third column shows the analyses in which the specific residue is ranked in top 20.

Residue No.	Residue	Analysis
42	M	seq,str,dyn
111	Y	seq,str,dyn
100	Y	seq,str
153	F	seq,dyn
4	L	str,dyn
61	I	str,dyn
92	M	str,dyn
93	V	str,dyn
112	L	str,dyn
113	T	str,dyn
154	E	str,dyn
11	D	seq
18	N	seq
21	P	seq
22	W	seq
23	N	seq
27	D	seq
32	K	seq
45	H	seq
52	R	seq
53	P	seq
55	P	seq
59	N	seq
81	A	seq
121	G	seq
125	F	seq
133	W	seq
3	S	str

Residue No.	Residue	Analysis
8	L	str
30	W	str
35	T	str
38	K	str
43	G	str
94	I	str
95	G	str
96	G	str
99	V	str
5	I	dyn
6	A	dyn
40	V	dyn
41	I	dyn
60	I	dyn
109	K	dyn
110	L	dyn
114	H	dyn
155	I	dyn
156	L	dyn

Table C.3: Top 20 residues identified for DHFR using each of the approaches -sequence, structure and dynamics are given in the table. Third column shows the analyses in which the specific residue is ranked in top 20.

Appendix D

Deep2Full: Evaluating Strategies for Selecting the Minimal Mutational Experiments for Optimal Computational Predictions of Deep Mutational Scan Outcomes

	RMSD					Pearson correlation				
Scan / Protein	Random 15%	Random 25%	Random 50%	Random 85%	SNS-Random 25%	Random 15%	Random 25%	Random 50%	Random 85%	SNS-Random 25%
β-lactamase	0.67	0.64	0.57	0.54	0.65	0.81	0.83	0.87	0.89	0.83
APH(3')-II	1.11	1.12	1.05	0.98	1.18	0.69	0.68	0.72	0.78	0.68
Hsp90	0.22	0.20	0.19	0.17	0.21	0.72	0.77	0.82	0.85	0.75
MAPK1	0.41	0.40	0.35	0.33	0.42	0.62	0.63	0.74	0.77	0.63
UBE2I	0.33	0.30	0.28	0.27	0.30	0.52	0.59	0.66	0.67	0.61
TPK1	0.39	0.39	0.38	0.35	0.42	0.24	0.23	0.26	0.42	0.24

Table D.1: RMSD and Pearson correlation for the test set of scans varying the number of training data points.

	RMSD					Pearson correlation				
Scan / Protein	ANH Scan	Random 15%	Position range scan	WT residue type scan	SASA range scan	ANH Scan	Random 15%	Position range scan	WT residue type scan	SASA range scan
β-lactamase	0.71	0.67	0.92	0.90	1.03	0.80	0.81	0.65	0.67	0.53
APH(3')-II	1.13	1.11	1.32	1.31	1.39	0.67	0.69	0.52	0.54	0.49
Hsp90	0.22	0.22	0.39	0.31	0.33	0.75	0.72	0.30	0.37	0.50
MAPK1	0.42	0.41	0.57	0.52	0.55	0.62	0.62	0.31	0.39	0.33
UBE2I	0.31	0.33	0.46	0.40	0.41	0.56	0.52	0.20	0.32	0.31
TPK1	0.39	0.39	0.45	0.40	0.43	0.25	0.24	0.13	0.19	0.10

Table D.2: RMSD and Pearson correlation for the test set of the 15% scans.

Variable	Pearson correlation with EVmutation
Conservation	-0.49
SASA	0.39
Contacts	-0.36
Average commutetime	0.34
Average co-evolution	-0.32
Closeness centrality	-0.27
Eigenvector centrality	-0.25
Degree centrality	-0.23

Table D.3: Table showing the Pearson correlation of different variables that was considered in our study as inputs for the neural network with the EVmutation score. Negative values indicate anti-correlation.

	Pearson correlation		RMSD	
	Random 85%	Envision	Random 85%	Roth et al.
β-lactamase	0.89	0.85	0.54	-
APH(3')-II	0.78	0.84	0.98	-
Hsp90	0.85	0.76	0.17	-
UBE2I	0.67	-	0.27	0.24
TPK1	0.42	-	0.35	0.34

Table D.4: Comparison of prediction quality of Deep2Full with other methods which used partial deep scan data to complete the map. For Envision the Pearson correlation for the test set of individual protein models developed by training on 80% of deep mutational scan data was obtained from Figure 2 of *Gray et al.*[1].

	Spearman correlation								
Protein	Deep-Sequence	EVmutation	SNAP2	Envision	ANH scan	Random 15%	Random 25%	Random 50%	Random 85%
β-lactamase	0.78	0.72	0.71	0.74*	0.80	0.81	0.83	0.86	0.88
APH(3')-II	0.59	0.54	0.49	0.64*	0.67	0.68	0.67	0.72	0.77
Hsp90	0.53	0.49	0.43	0.31*	0.53	0.56	0.59	0.65	0.70
MAPK1	-0.24	-0.25	0.30	-0.44	0.60	0.59	0.60	0.71	0.75
UBE2I	0.55	0.51	-0.51	0.09	0.56	0.52	0.59	0.65	0.66
TPK1	0.26	0.25	-0.22	0.27	0.25	0.24	0.24	0.26	0.42

Table D.5: Comparison of prediction quality of our models with that of existing methods which do not use partial data for generating the model. For DeepSequence[2] and EVmutation[3], the data was taken from the supplementary information of *Riesselman et al.*[2]. *Extracted from the supplementary figure 8 on Leave-One-Protein-Out analysis of *Gray et al.*[1].

Scan / Protein	Random 85%	Random 50%	Random 25%	Random 15%	ANH scan	Position range scan	WT residue type scan	SASA range scan
blact	41	31	30	15	13	20	20	20
agk	26	33	28	15	15	20	12	20
hsp90	42	30	12	12	11	20	9	20
mapk1	35	21	28	19	20	17	15	20
ube2i	16	12	40	25	23	20	14	18
tpk1	40	40	36	15	20	20	15	14

Table D.6: Optimal number of hidden neurons for all proteins and scans.

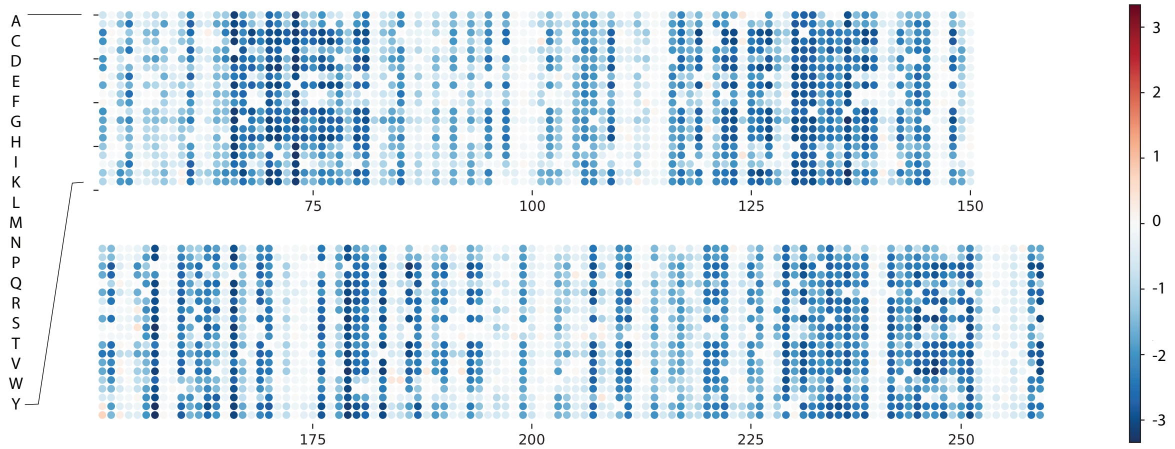

Figure D.1: Computational fitness map: Computational predictions of the quantitative gain or loss in fitness of *E. coli* resulting from amino acid changes in β-lactamase when challenged with 2500 μg/ml concentration of ampicillin. The panel includes the training, validation and test sets. In this specific case the training and validation sets add up to 85% of the mutational scan data.

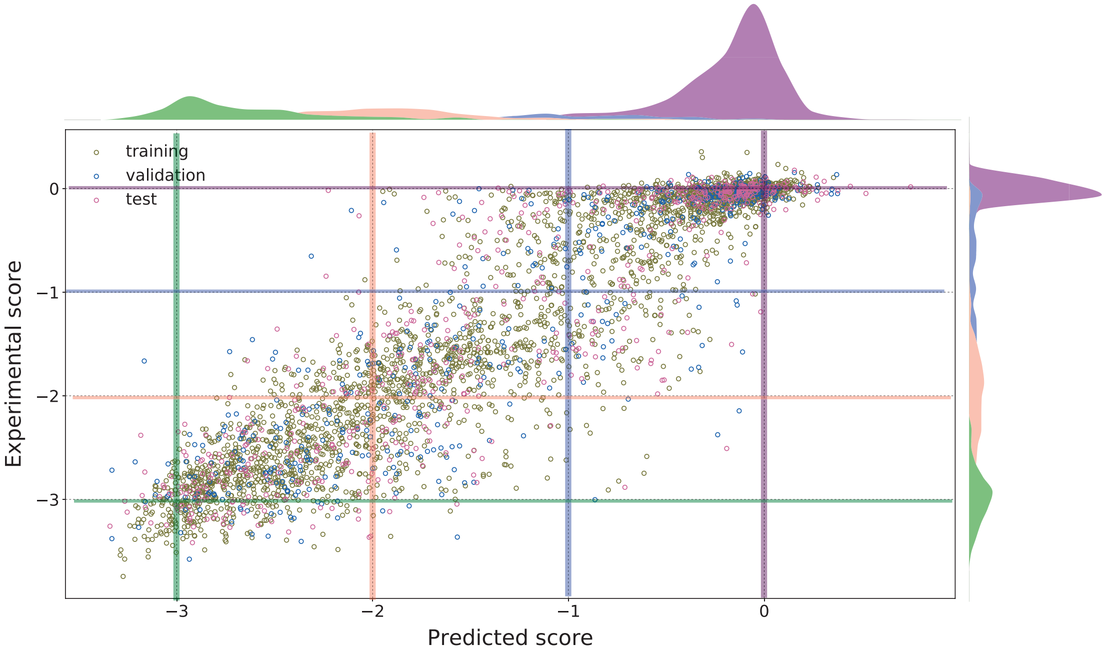

Figure D.2: Distribution of predictions over the range of fitness: Quality of predictions when the model was trained with 85% data is shown at different sections of the data. Distributions were constructed by taking data which are 0.3 units wide (0.15 to -0.15, -1.15 to -0.85, -2.15 to -1.85 and -3.15 to -2.85). The predictions made around fitness score 0 or predicted as neutral are more reliable as the spread in the experimental data corresponding to these predictions is lower.

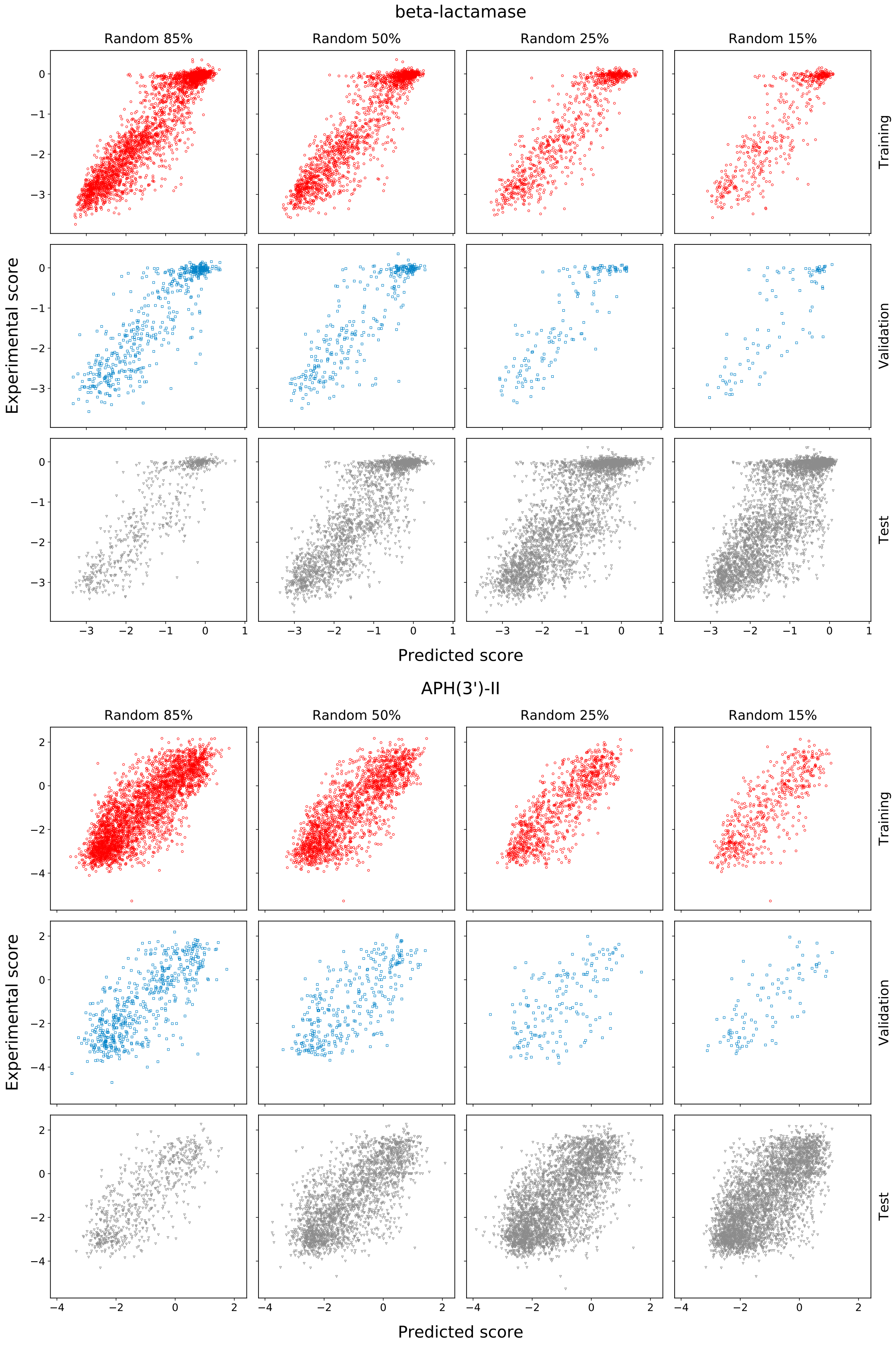

Figure D.3: Results from training the model with systematically reduced data sets: Fitness scores obtained computationally by training the models on decreasing sizes of data sets. Training, validation and test sets for β-lactamase and APH(3′)-II are shown. The RMSD and Pearson correlation for the test set for all proteins and for all scans are given in Table D.1.

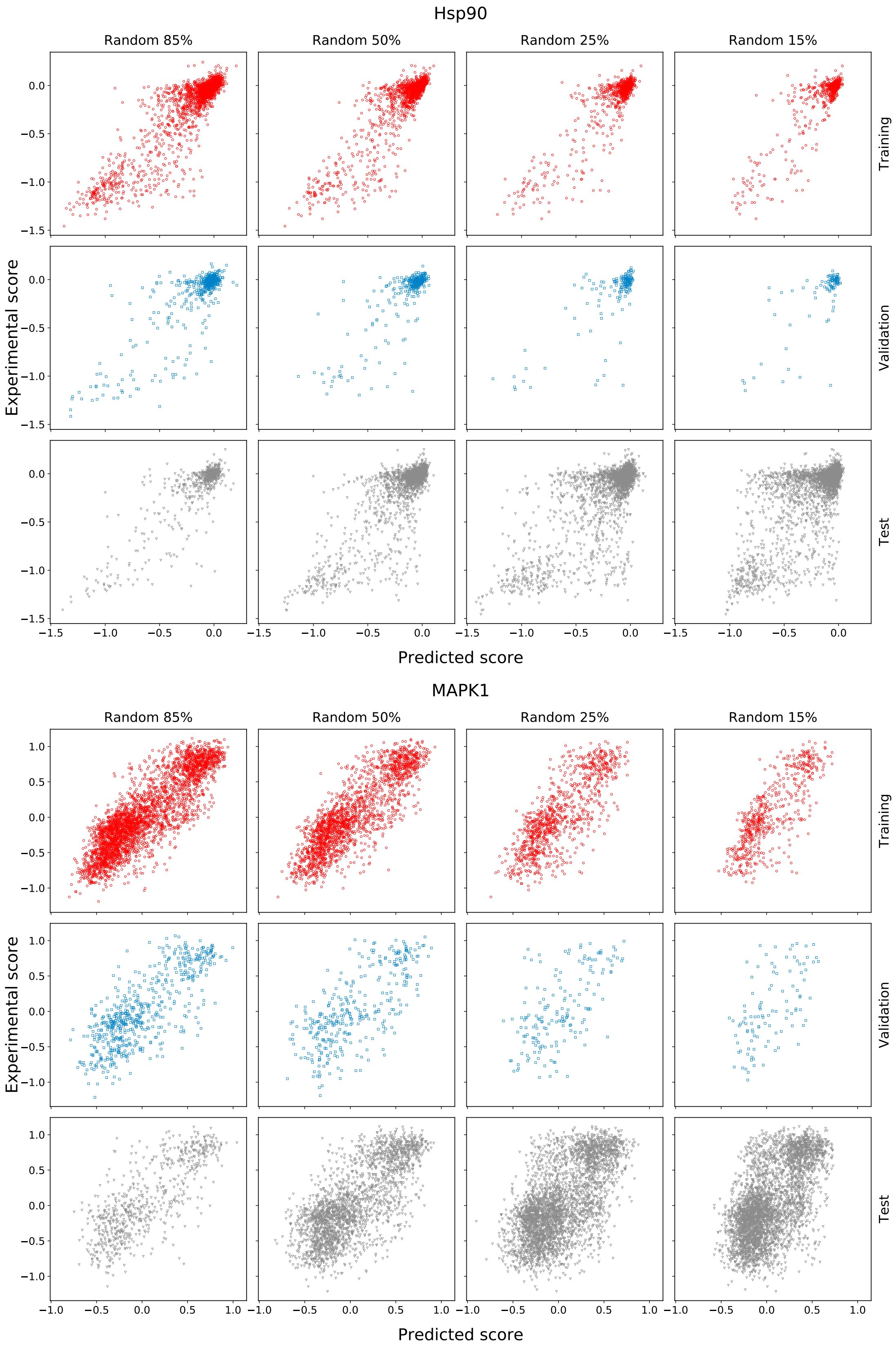

Figure D.4: Comparison of the predicted scores using models developed by training on datasets of varying sizes with the experimental fitness scores for Hsp90 and MAPK1.

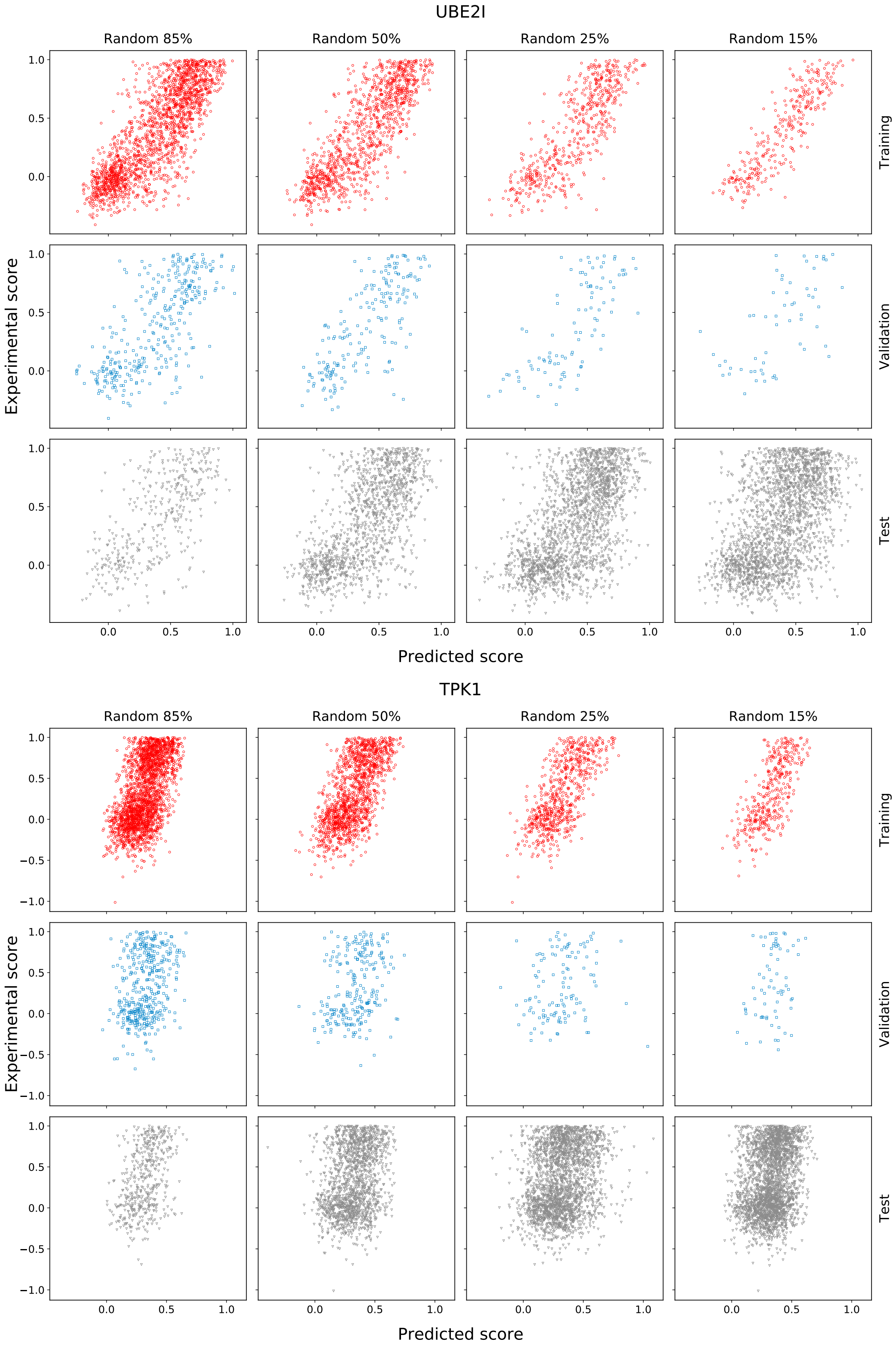

Figure D.5: Fitness scores obtained computationally by training the models on decreasing sizes of data sets. Training, validation and test sets for UBE2I and TPK1 are shown.

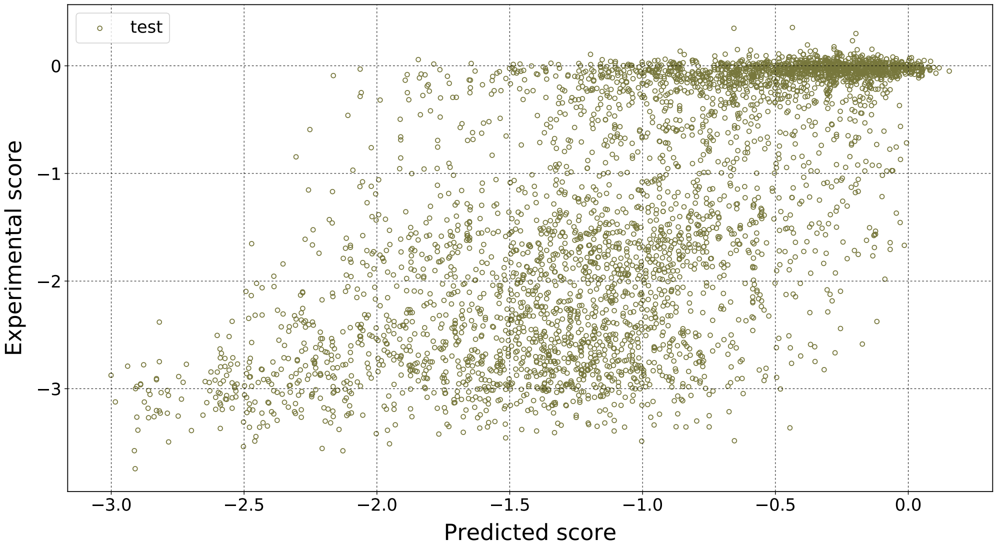

Figure D.6: Insufficiency of alanine scan for training the model: The experimental fitness data on substitution at every amino acid position of β-lactamase with alanine was used for training the model. This strategy which only used 5% of the full mutational data for predicting the fitness of all other 19 mutational scans did not give good predictions. We did not use alanine scan for any further analysis.

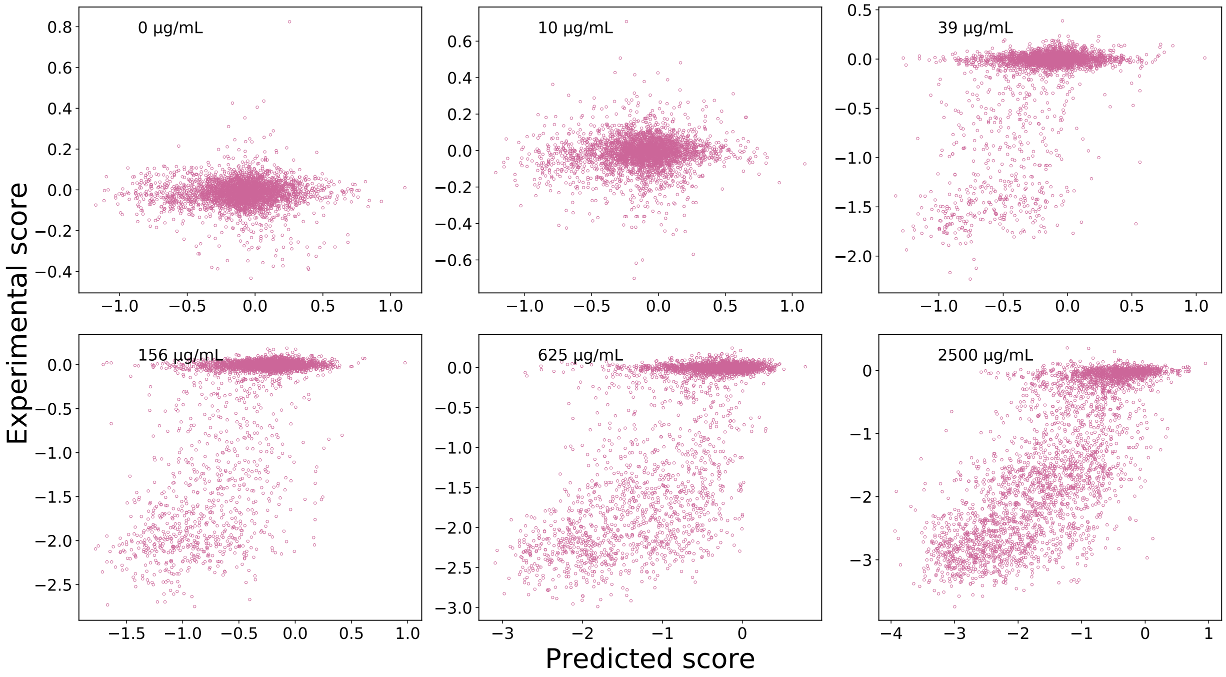

Figure D.10: Results from augmenting the data with transverse assay conditions: For the same amino acid substitutions we included the fitness data obtained at different antibiotic concentrations: 0, 10, 39, 156, 625 and 2500 μg/mL in the training set. Predictions for the remaining 85% of the mutations that were not used in the training are shown and augmenting do not improve the results.

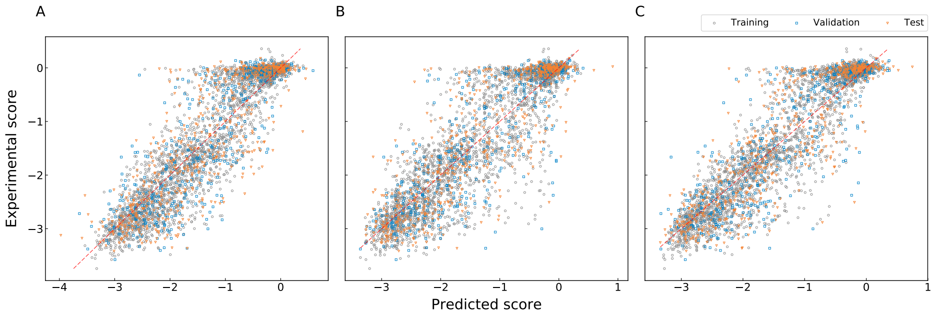

Figure D.11: Feature selection: Predictions from models with reduced number of variables. (A) Using the variables impact, average correlation, average commute time, contacts, BLOSUM and hydrophobicity of mutant which were chosen based on the variable impact analysis (B) Using variables chosen based on the pearson correlation between the input variables and fitness: conservation, average correlation, average commute time, contacts, BLOSUM, SASA, PSSM score for wild type amino acid (C) Using all 17 variables. The models with fewer variables have comparable predictive abilities as that the one using all variables. The adjusted R^2 values for the test set are 0.74, and 0.74 and 0.78 respectively for A, B and C. The variables average correlation, average commute time, contacts and Blosum are common in A and B.

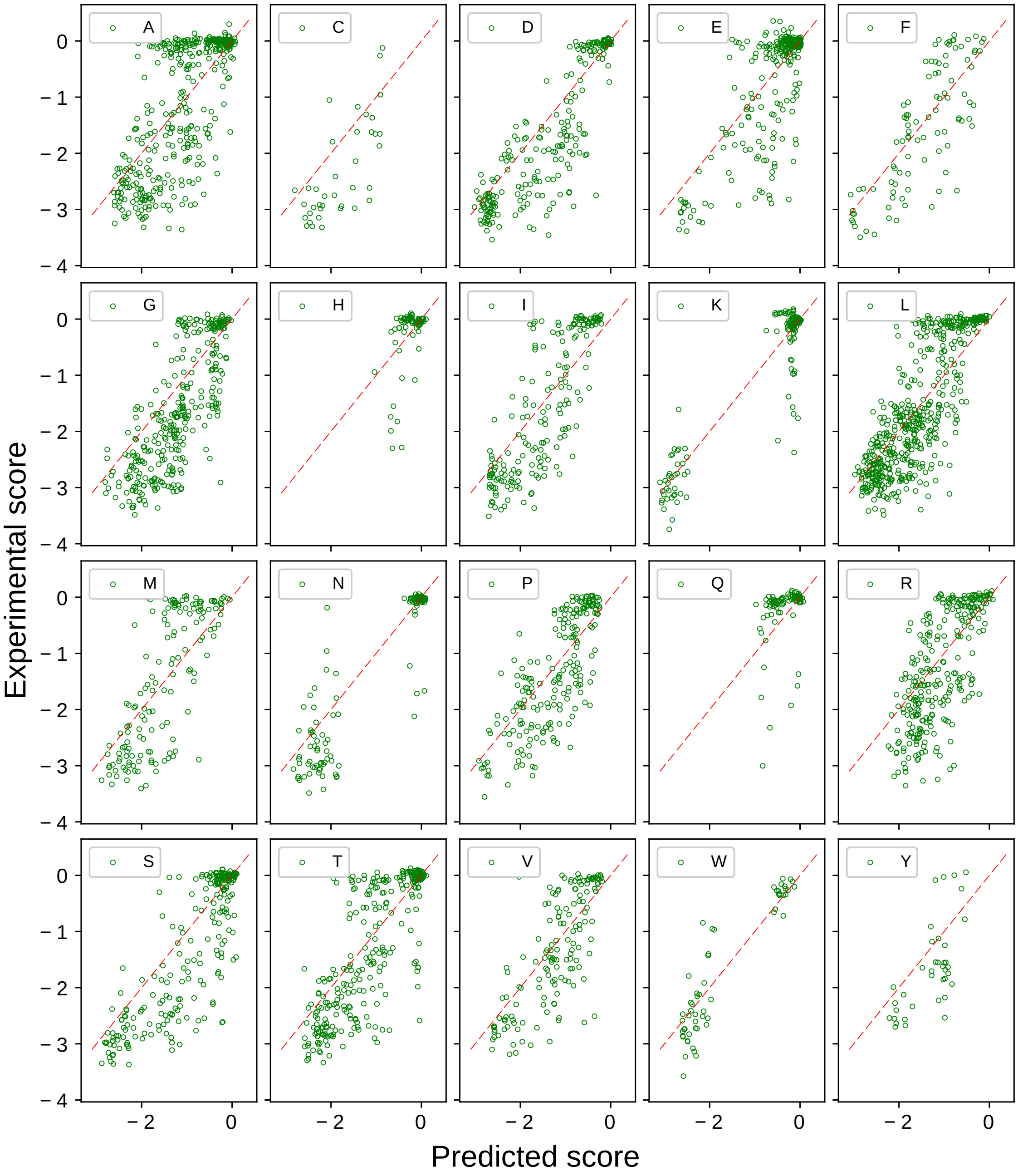

Figure D.12: Prediction quality relative to the wild type amino acid: Predictions from ANH were also analyzed by classifying them according to the wild type amino acid *which gets mutated*. The analysis shows that there is variability in the predictability of substitution of different amino acids such as asparagine (N) and tryptophan (W) having high prediction quality and histidine (H) and glutamine (Q) relatively poor. The dashed red lines are guidelines with slope 1.

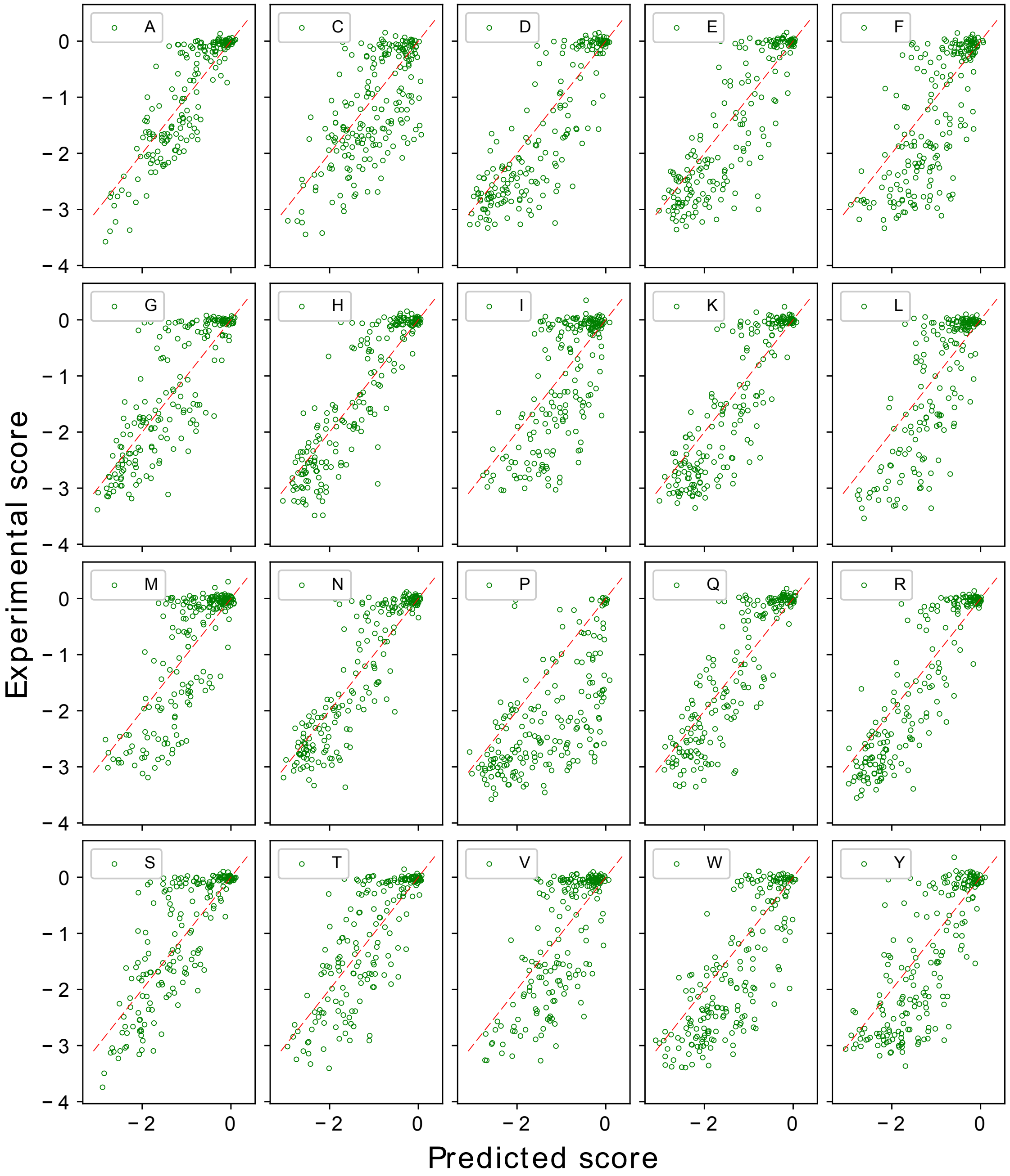

Figure D.13: Quality of output relative to the substituted amino acid: Predictions from ANH were analysed by classifying them according to the amino acid to which *mutation is performed*. It can be seen that the predictability of all amino acids is comparable. However, the predictions to alanine (A), asparagine (N), histidine (N) are notably better because of the training set that was used. The dashed red lines are guidelines with slope 1.

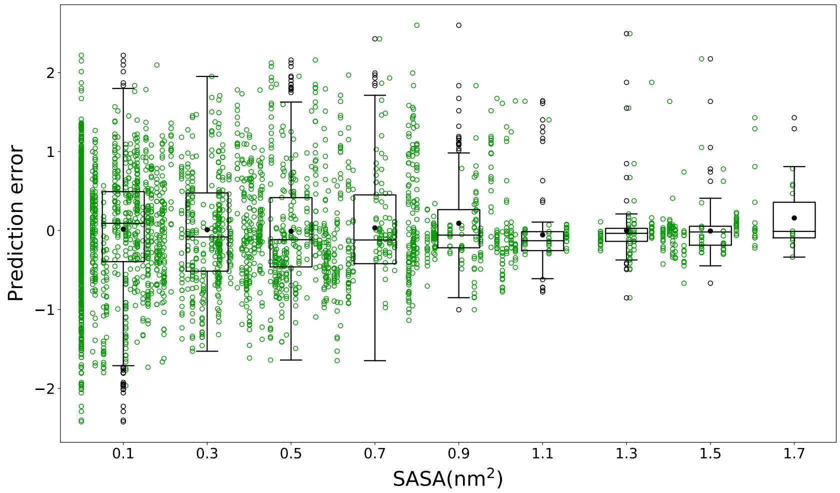

Figure D.14: Distribution of error along SASA: Prediction error (Predicted fitness - experimental fitness) for the test set of random 15% scan for β-lactamase as a function of solvent exposure. The box plot shows the distribution of errors at different values of SASA. At higher values of SASA, error is lower.

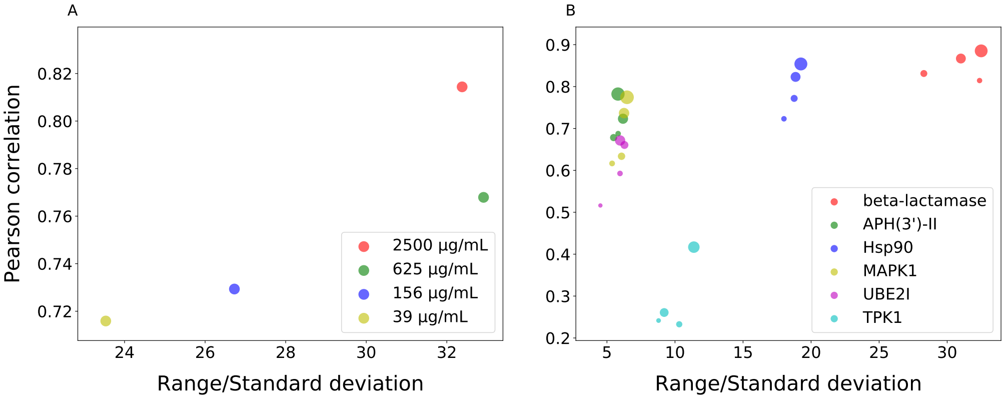

Figure D.15: Output quality versus input quality: Quality of data used in training defined as the ratio of the range of mutational effect scores in the training set and the standard deviation of the mode corresponding to the neutral substitutions with the prediction quality shown in Pearson correlation between the predictions and measured fitness for the test set was used as a measure of the quality of output. Quality of input vs and quality of output is compared for (A) models developed for mutational effect scores measured for β-lactamase under different concentrations of ampicillin (B) random scans performed for the six proteins. Comparison in the same system has a trend, while no clean trend could be seen in the comparisons across different protein and scans.

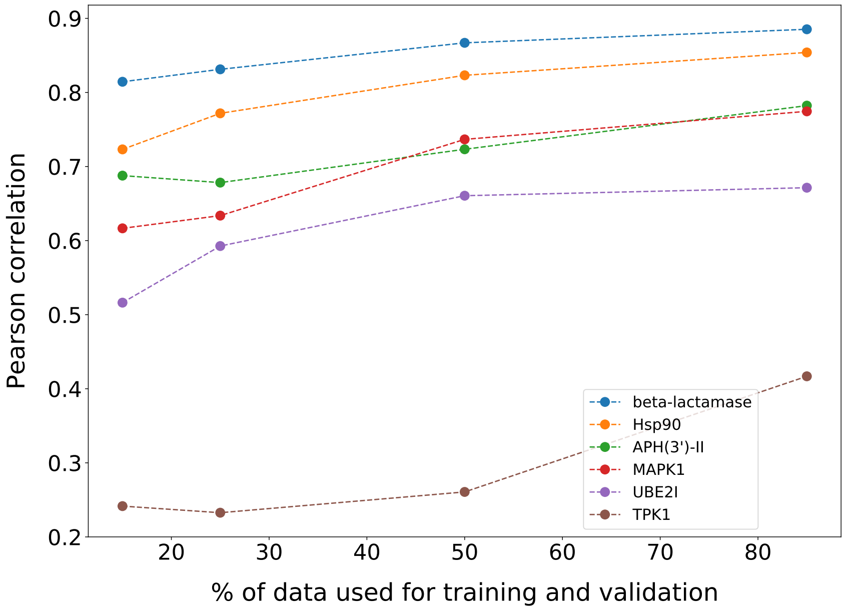

Figure D.16: Quality of predictions using different random scans: The quality of predictions as seen from the Pearson correlation between predicted and experimental data in general improve for all proteins. This improvement slows down beyond 50% data usage except for TPK1.

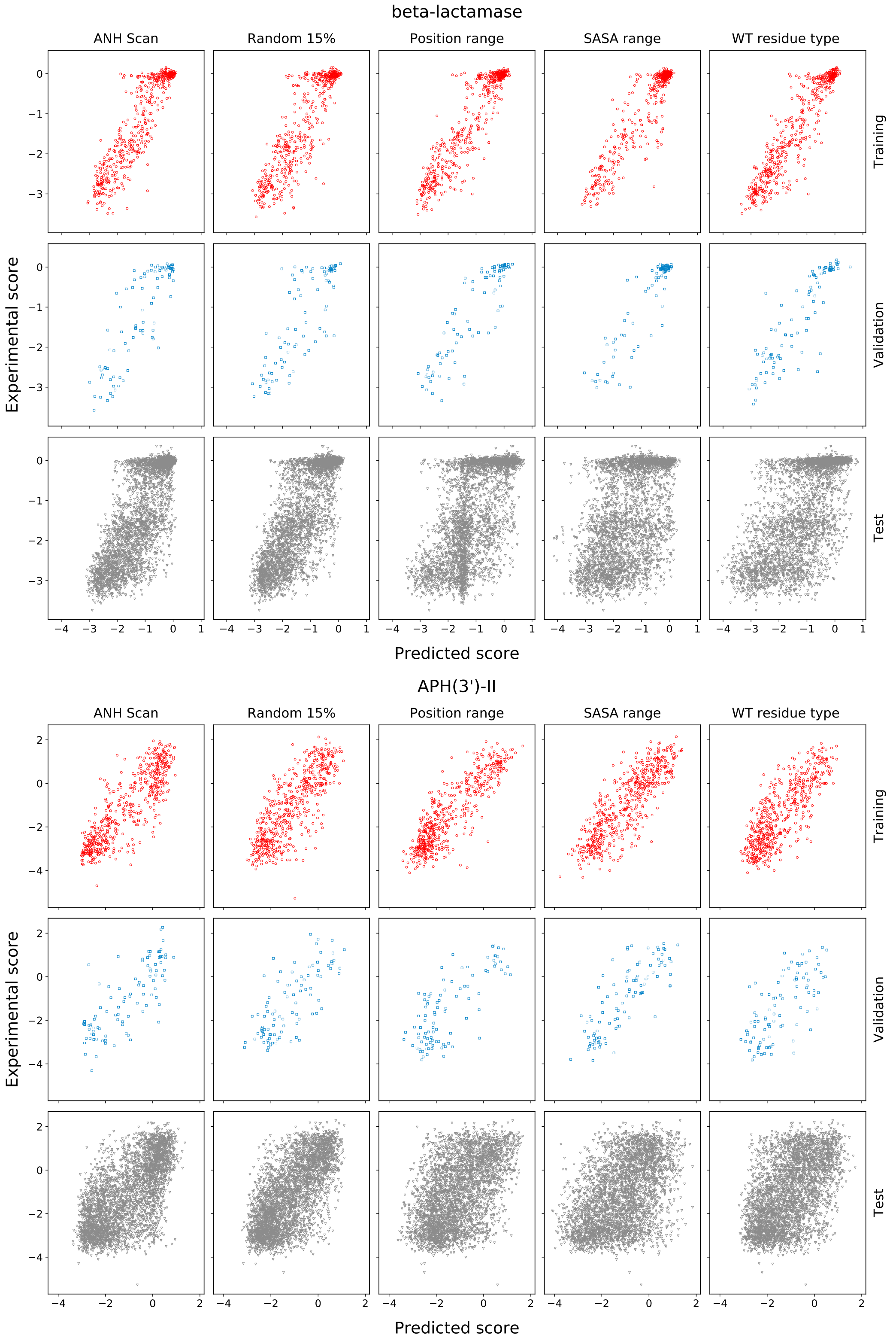

Figure D.7: Results from training the model on mutations chosen with different strategies: Fitness scores obtained computationally by training the models on 15% of data chosen in different ways. The training, validation and test sets for β-lactamase and APH(3′)-II are shown. The RMSD and Pearson correlation for the test set for all proteins and for all scans are given in Table D.2.

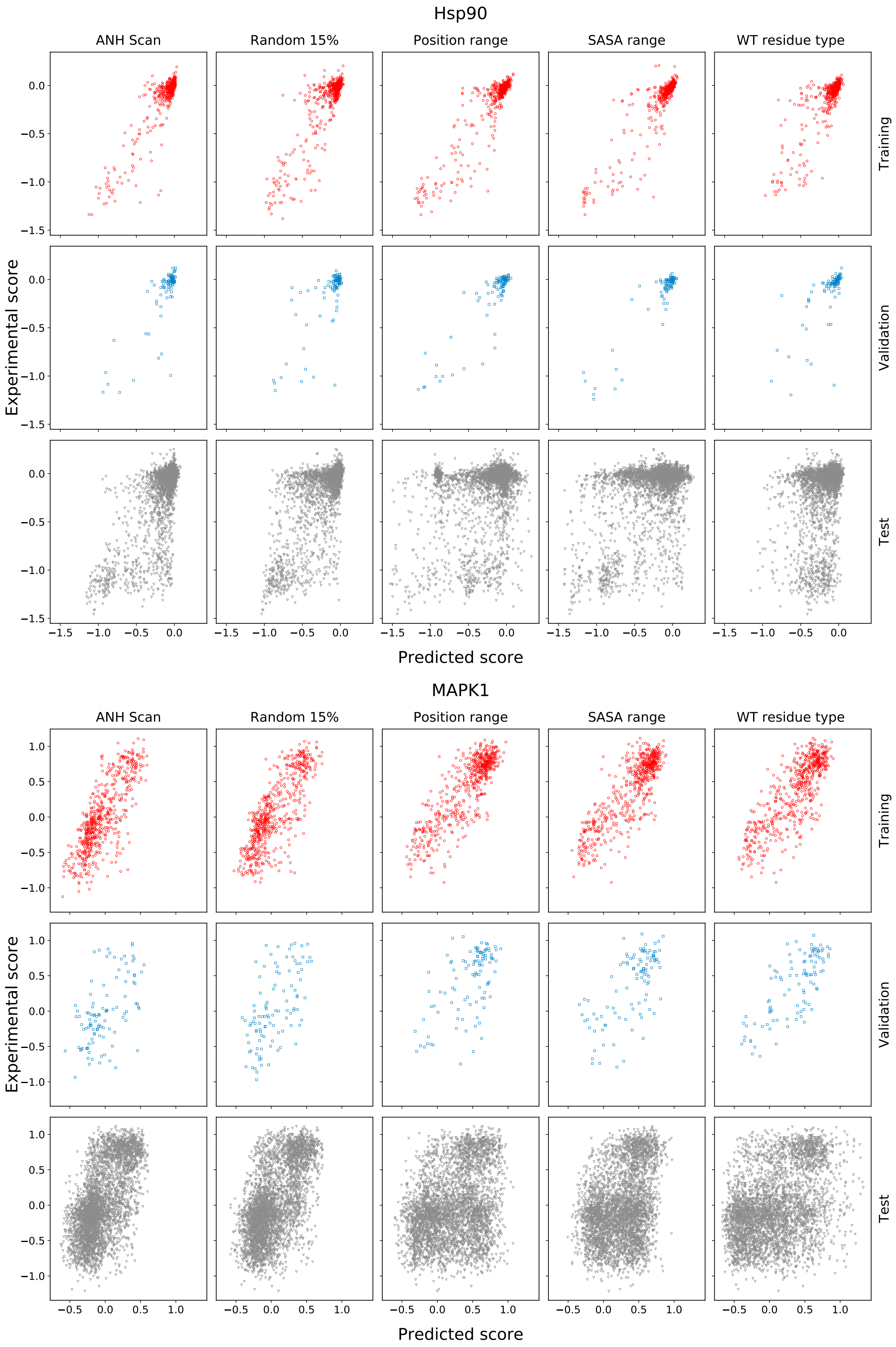

Figure D.8: Computational predictions of the models trained with 15% of data chosen with different strategies for Hsp90 and MAPK1.

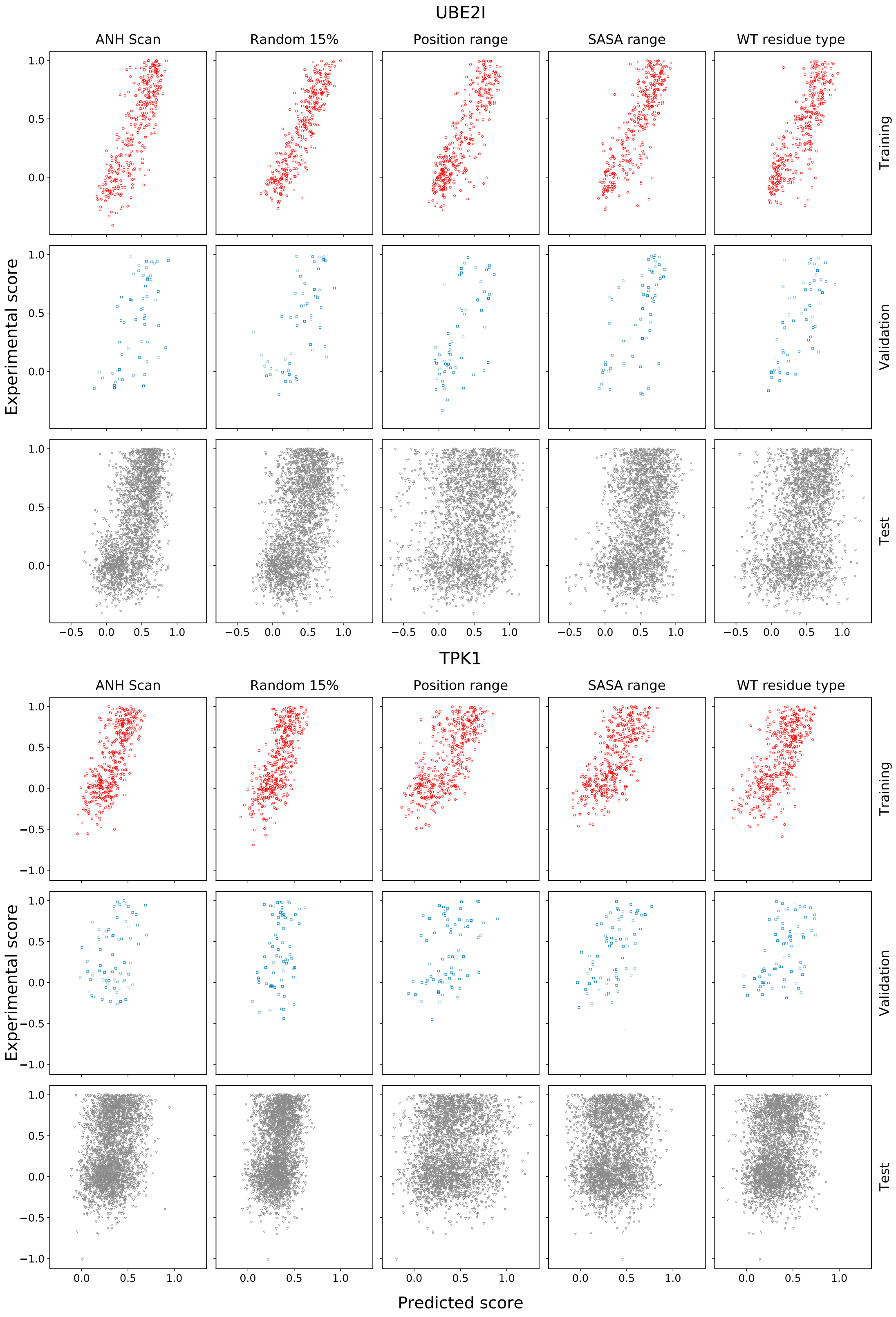

Figure D.9: Computational predictions of the models trained with 15% of data chosen with different strategies for UBE2I and TPK1

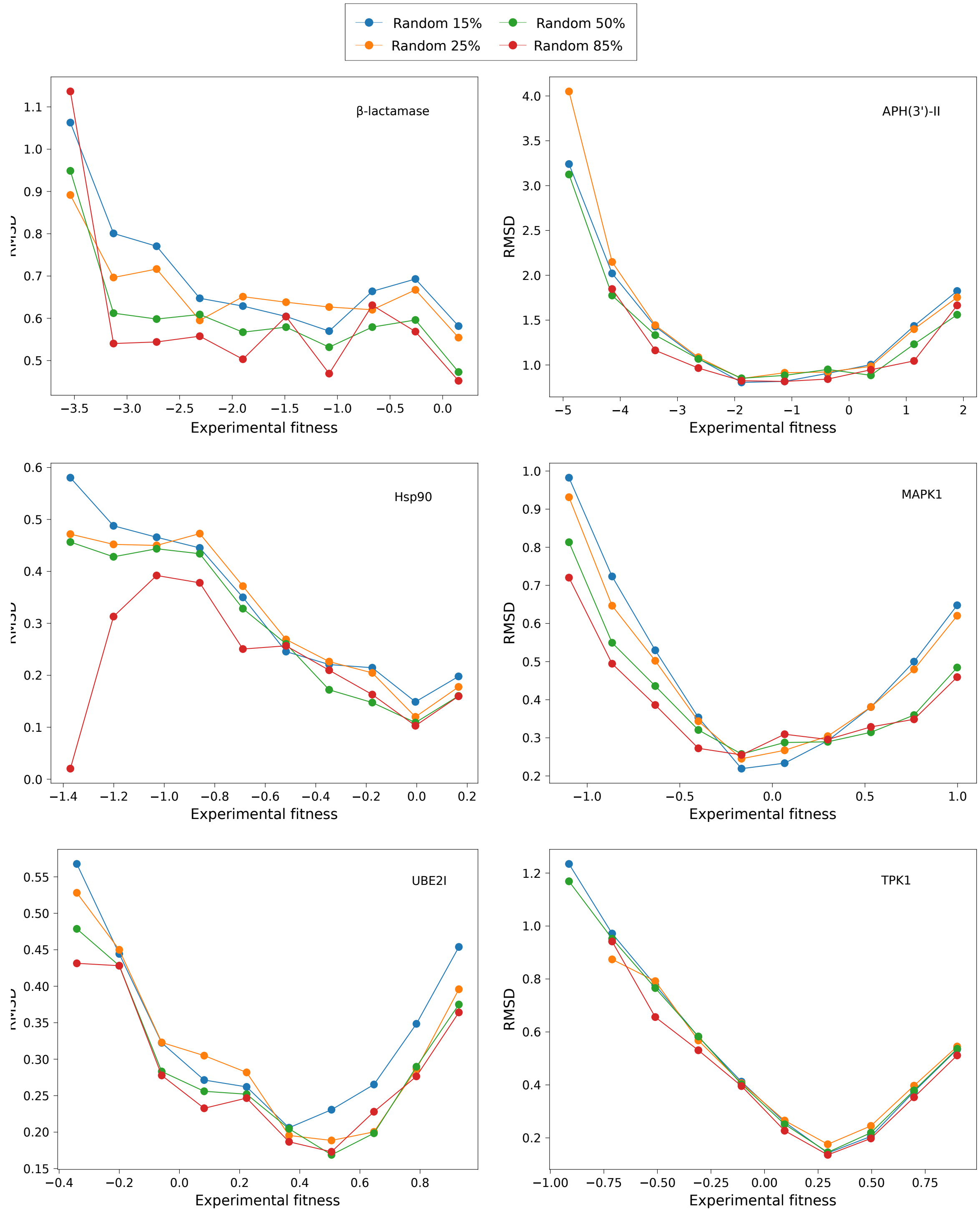

Figure D.17: Distribution of error over the range of fitness: RMSD of test set calculated across the range of experimental fitness for all the six proteins. The experimentally observed fitness was divided into 10 bins in each of the cases and RMSD was quantified for each bin. As it can be seen, errors are fairly systematic, which suggests that the RMSD relative to the expected values is fairly a constant for several of the cases.

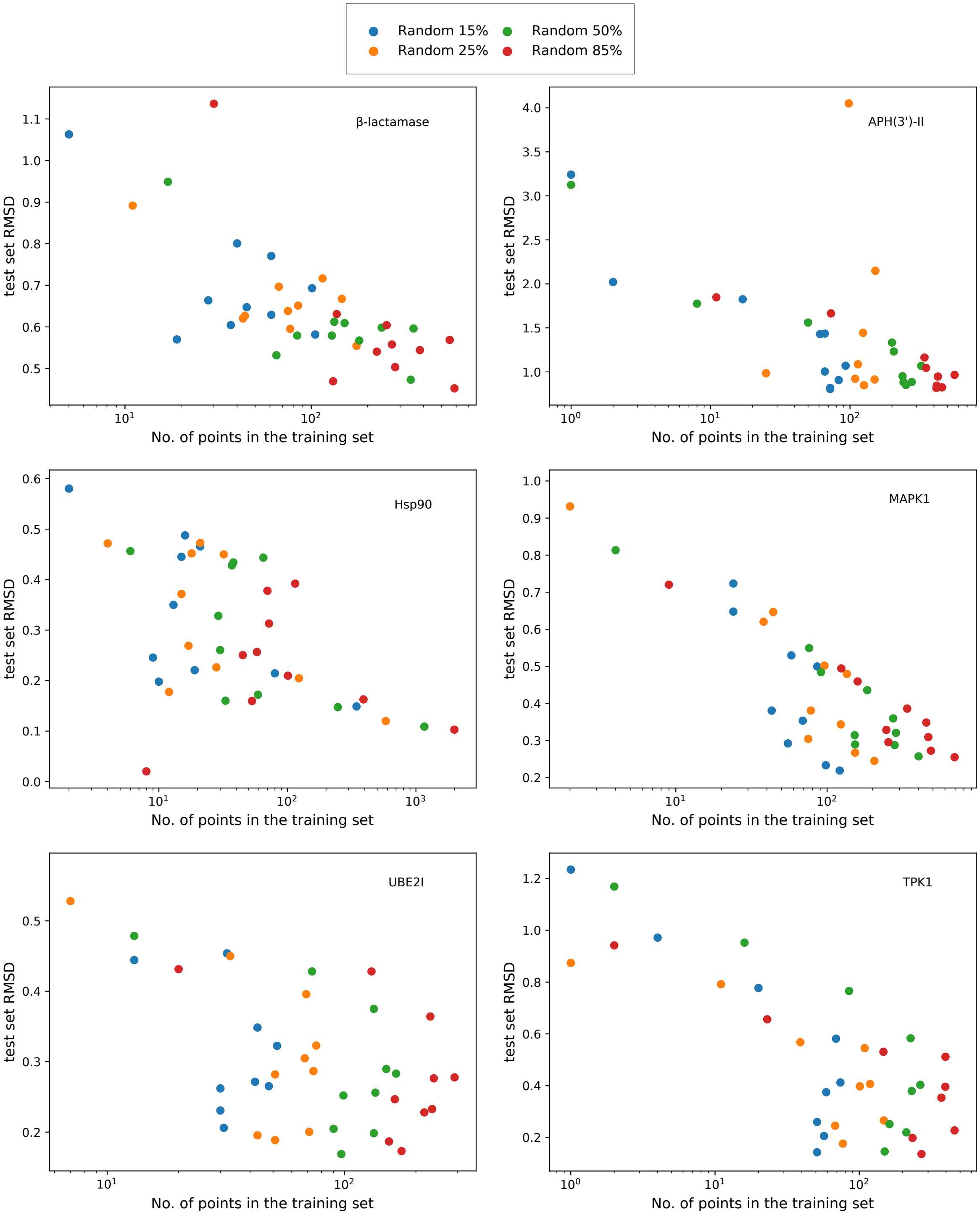

Figure D.18: Reduction of error with the increasing number of training points per bin: RMSD of test set calculated for different ranges of the experimental fitness scores versus the number of points in the training set in each bin. A power-law behavior emerges. It can also be seen that having more than 100 points for training from a bin does not improve the predictions significantly, which suggests an approximate 100×10 bins = 1000 data points for training.

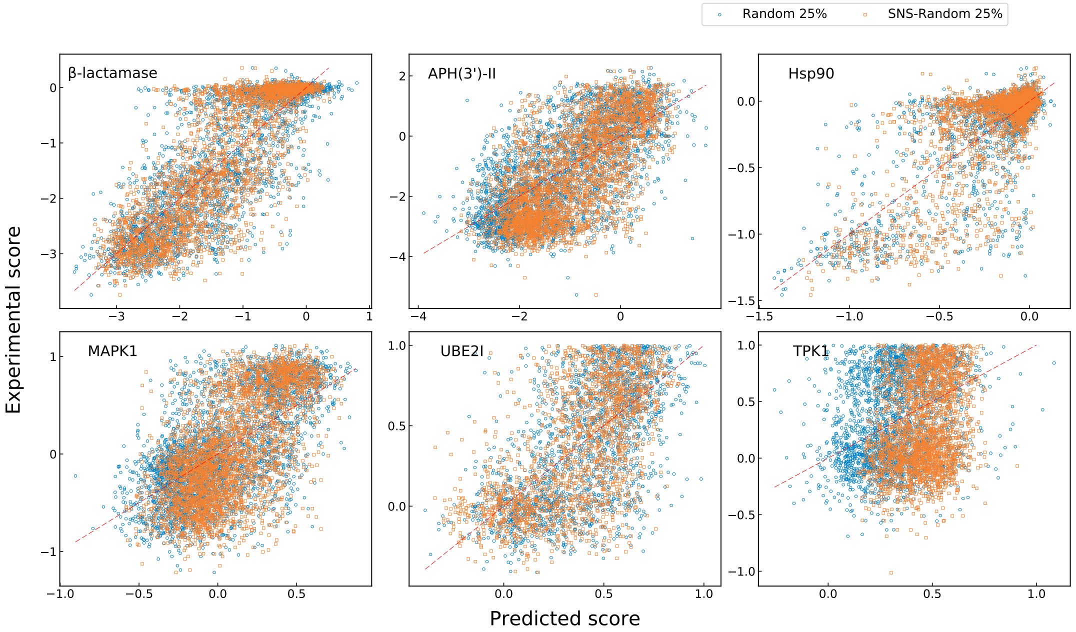

Figure D.19: Comparing the results of randomly choosing from SNS versus randomly choosing from all possible substitutions: Comparison of quality of predictions by models generated by training on randomly selected variants from the complete data and randomly selected mutations achievable through single nucleotide substitutions (SNS). 25% of data was used for training and validation. The RMSD and Pearson correlation for the test set for both scans and for all proteins are given in Table D.1. It can be seen that the performance of both models are comparable.

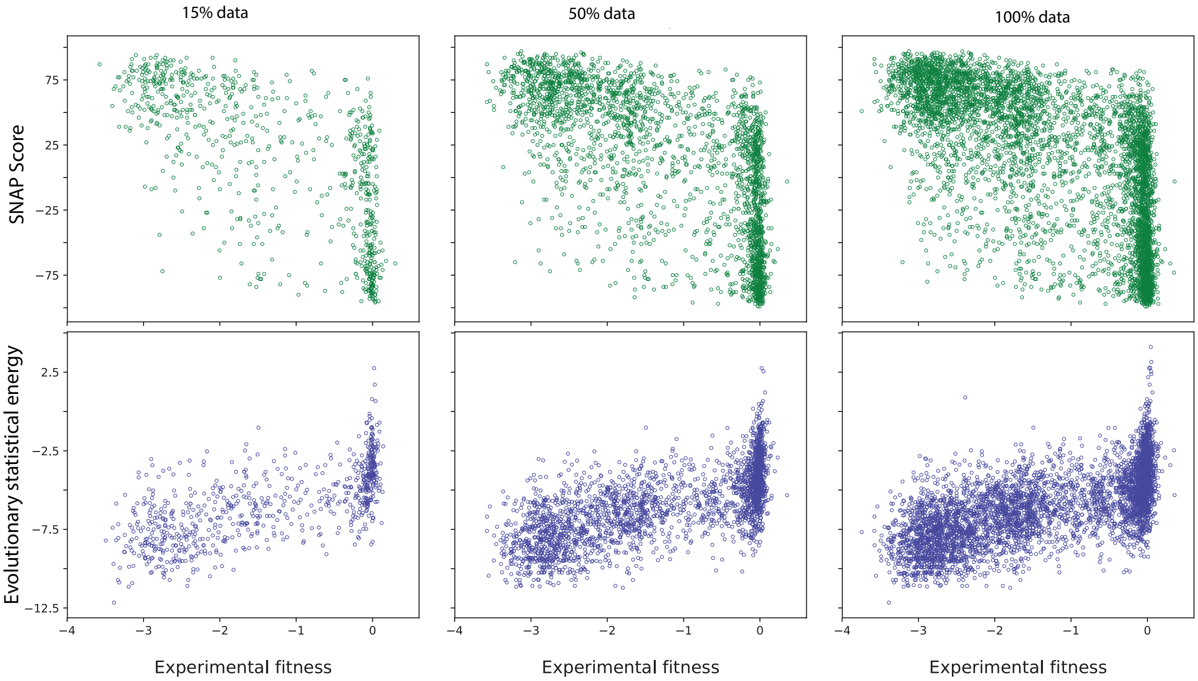

Figure D.20: Comparison of mutational effect score predictions from unsupervised methods with the experimental score: A comparison of how the scores from SNAP2 and Evolutionary statistical energy relate to the experimentally observed fitness of *E. coli* arising from β-lactamase mutations is shown in this figure. The scores are compared for the complete experimental data as well as by randomly selecting 15%, 50% of the data.

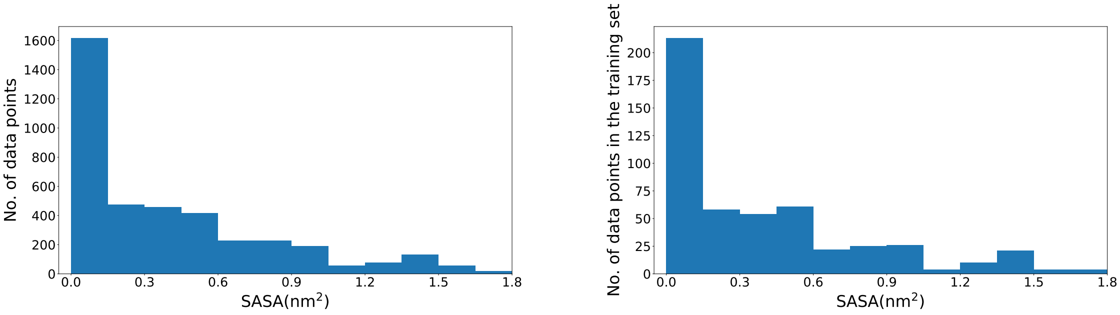

Figure D.21: Random scan is representative: Distribution of SASA for the complete data as well as the training set of random 15% scan for β-lactamase. The distributions are similar showing that the random choice of mutations is representative of the complete data set.

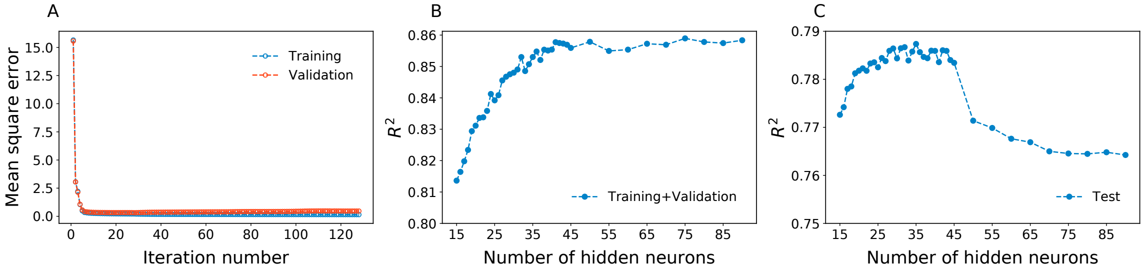

Figure D.22: Convergence of neural network model training and test: Results demonstrating how the choice of number of hidden neurons was made for β-lactamase for the random 85% scan. **(A)** For a given choice for the number of neurons in the hidden layer, 41 in this case, mean square error calculated for the training set as well as the validation set as the iteration proceeds. The quality of predictions for training data improves, and beyond the optimal number of iterations, the error of predictions for the validation set increases. To avoid this problem of overtraining early stopping criterion was used to terminate the iterations. When this worsening consistently occurs for 100 iterations, the training is terminated. The iteration at which the validation set has the lowest error is then chosen. **(B)** Similar calculation was repeated by changing the number of neurons, and the R^2 for training and validation was monitored as a function of the number of neurons. The decision about the optimal number of neurons was made considering the performance both in the training and the validation sets. In this case, 41 neurons in the hidden layer was considered optimal. **(C)** R^2 value for test set, obtained by performing the calculations at any given number of neurons is also shown. In this specific case, this graph is used an *a posteriori* justification for our choice of 41 neurons in the hidden layer. As it can be seen although with higher number of neurons the training and validation may saturate, the test set can worsen, and attention needs to be paid to avoid this problem by over training.

Bibliography

[1] Gray VE, Hause RJ, Luebeck J, Shendure J, Fowler DM. Quantitative missense variant effect prediction using large-scale mutagenesis data. Cell systems. 2018;6(1):116–124.

[2] Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. Nature Methods. 2018;15(10):816+. doi:10.1038/s41592-018-0138-4.

[3] Hopf TA, Ingraham JB, Poelwijk FJ, Scharfe CPI, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nature Biotechnology. 2017;35(2):128–135. doi:10.1038/nbt.3769.

Appendix E

Towards Developing Intuitive Rules for Protein Variant Effect Prediction Using Deep Mutational Scan Data

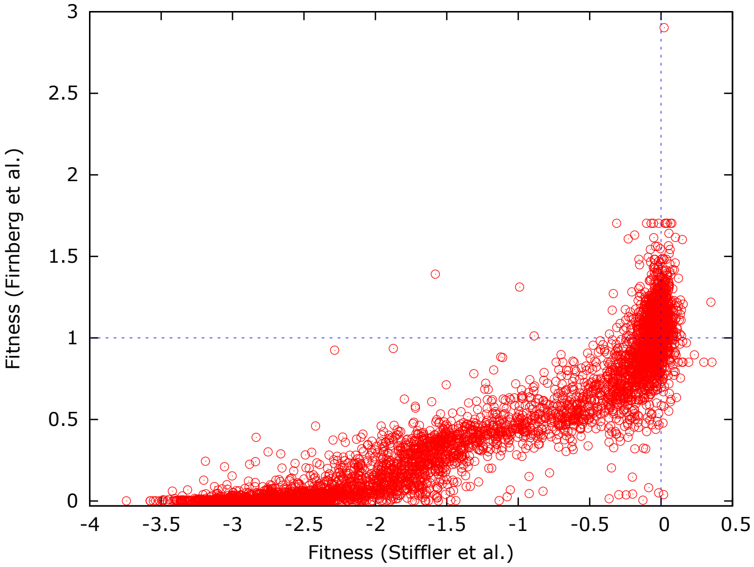

Figure E.1: A comparison of the results reported in two different deep mutational scan experiments where the mutations in β-lactamase and their fitness effects on *E. coli* were studied. The two dashed lines passing through 0 and 1 represent the wildtype fitness in the Stiffler *et al.*[1] and Firnberg *et al.*[2] experiments respectively.

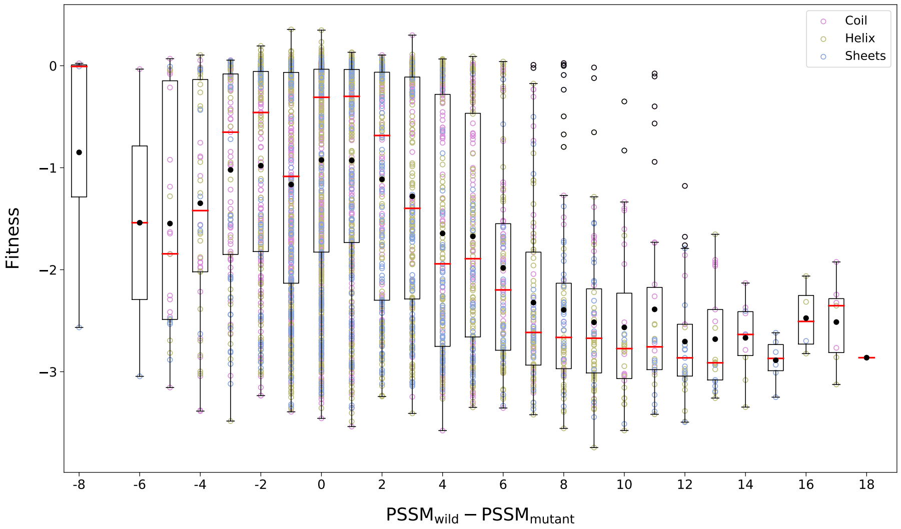

Figure E.2: Comparison of fitness with $\Delta PSSM$ ($= PSSM_{wildtype} - PSSM_{mutant}$) obtained from the Position Specific Scoring Matrix (PSSM). Conservation of amino acid is dependent only on the amino acid position and mainly capture average fitness effect upon substitution and cannot give information about specific substitutions being made. To capture the specific mutation information, we use the position specific scoring matrix (PSSM) as a reference, which has been developed by combining the knowledge about all possible substitutions. PSSM score was calculated using PSI-BLAST for the multiple sequence alignment (MSA) of β-lactamase. The fitness effect shows a dependence, albeit weak.

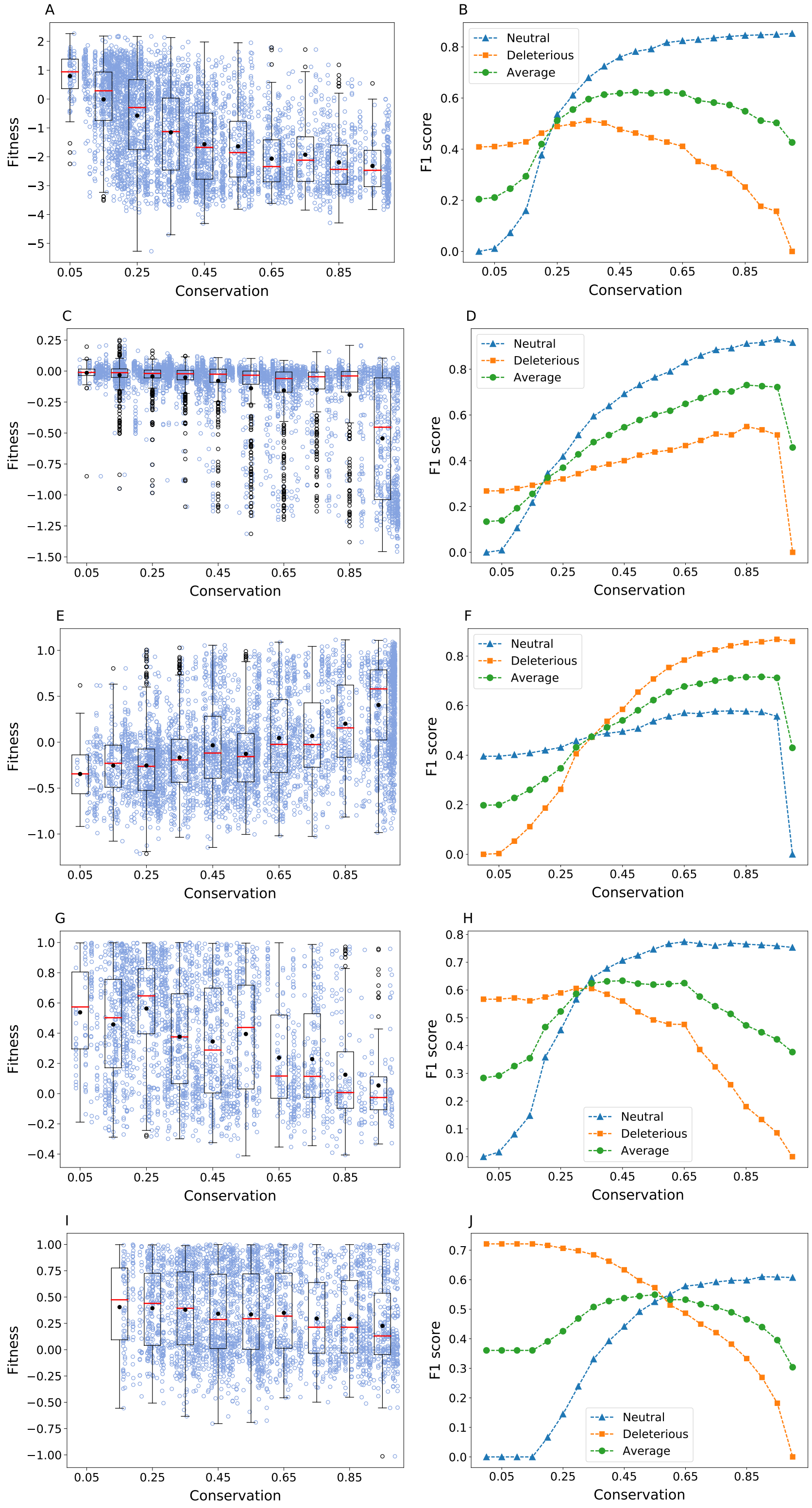

Figure E.3: Fitness dependence on the amino acid conservation at a position and the quality of classification quantified using F1 score as the conservation threshold for classification is varied are shown for the proteins APH(3')-II **(A,B)**, Hsp90 **(C,D)**, MAPK1 **(E,F)**, UBE2I **(G,H)** and TPK1 **(I,J)**. In the box plots, the black filled circle and the red line represent mean and median of the fitness respectively. The whiskers are plotted at 1.5 times the interquartile range and black open circles show the outliers.

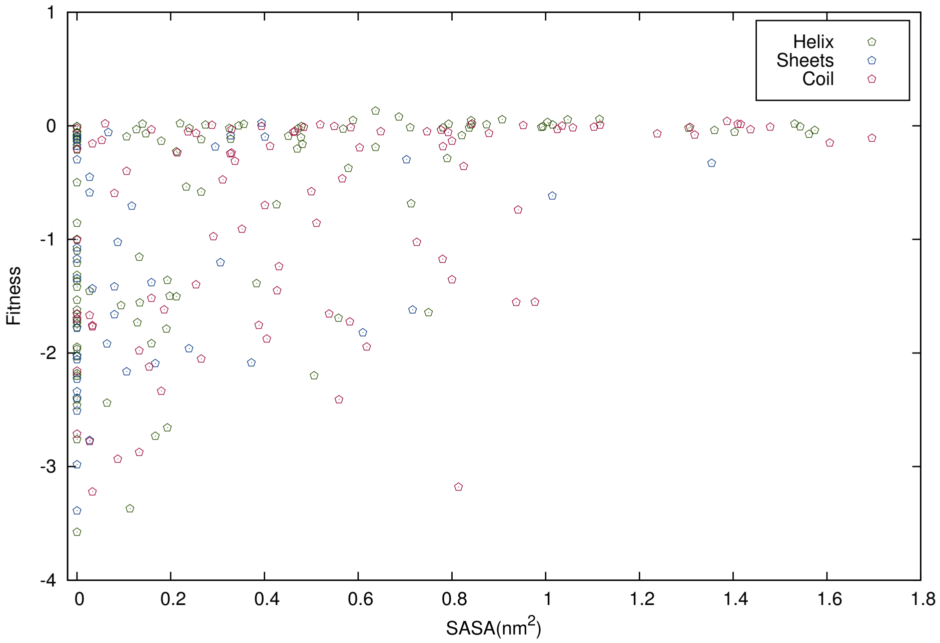

Figure E.4: For the sake of simpler interpretations the quantitative relation of fitness with SASA only for alanine substitutions is examined. The analysis reflects triangular pattern with solvent exposure.

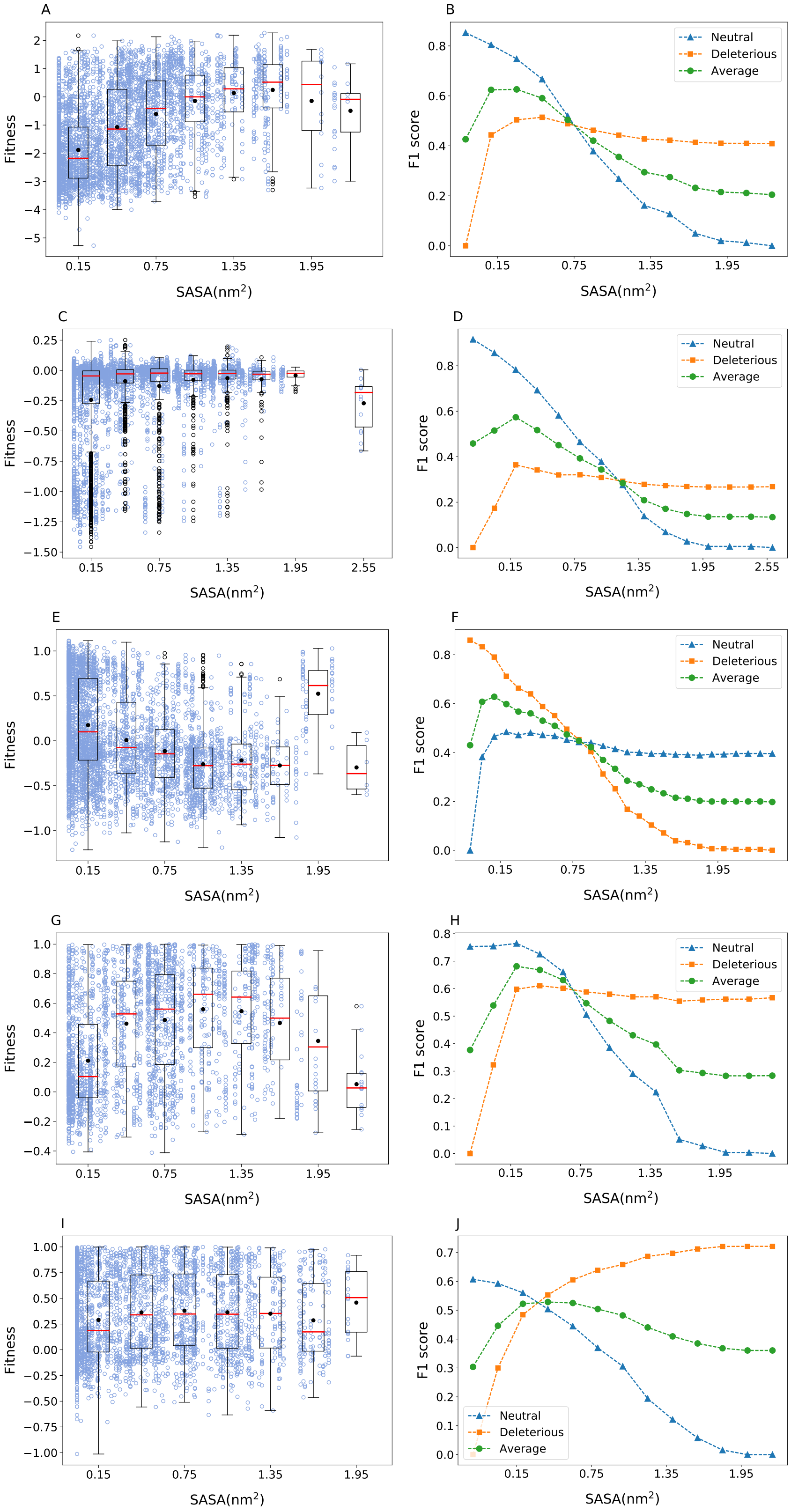

Figure E.5: Correlation of fitness with SASA and the F1 scores quantifying the quality of classification as the SASA threshold is varied for the proteins APH(3')-II **(A,B)**, Hsp90 **(C,D)**, MAPK1 **(E,F)**, UBE2I **(G,H)** and TPK1 **(I,J)**. Details about the box plot representation are given in Figure E.3.

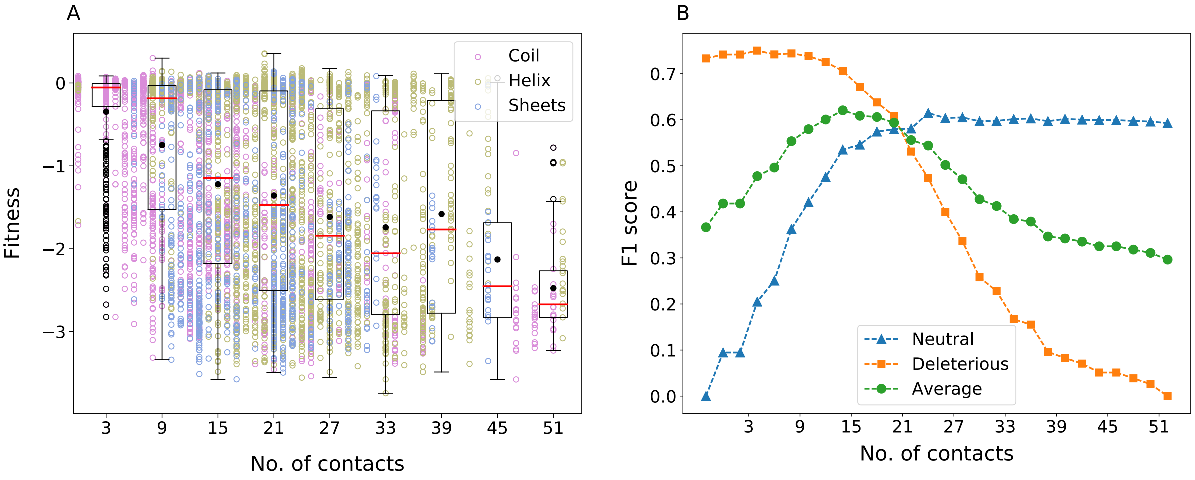

Figure E.6: A. The effect of packing was studied by plotting fitness relative to the number of contacts. No strong relation was observed. **B.** F1 score of neutral and deleterous classes and the average of both as the number of contacts threshold is changed. For details of box plot representation see Figure E.3.

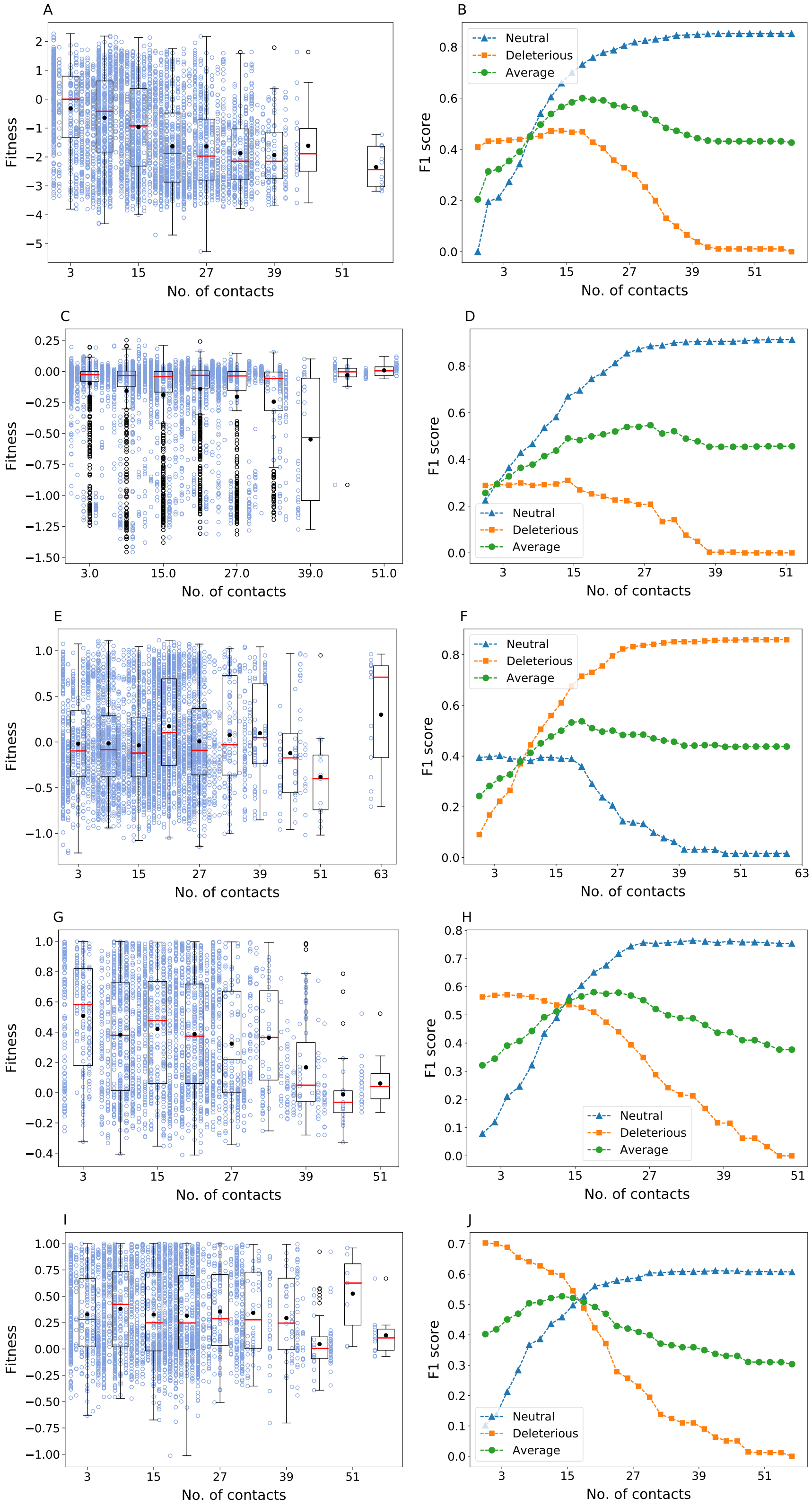

Figure E.7: Fitness scores with respect to number of contacts of the wild type amino acid and the F1 scores at different number of contacts thresholds chosen for classification. Data is shown for the proteins APH(3')-II **(A,B)**, Hsp90 **(C,D)**, MAPK1 **(E,F)**, UBE2I **(G,H)** and TPK1 **(I,J)**. (Box plot representation details are given in see Figure E.3.)

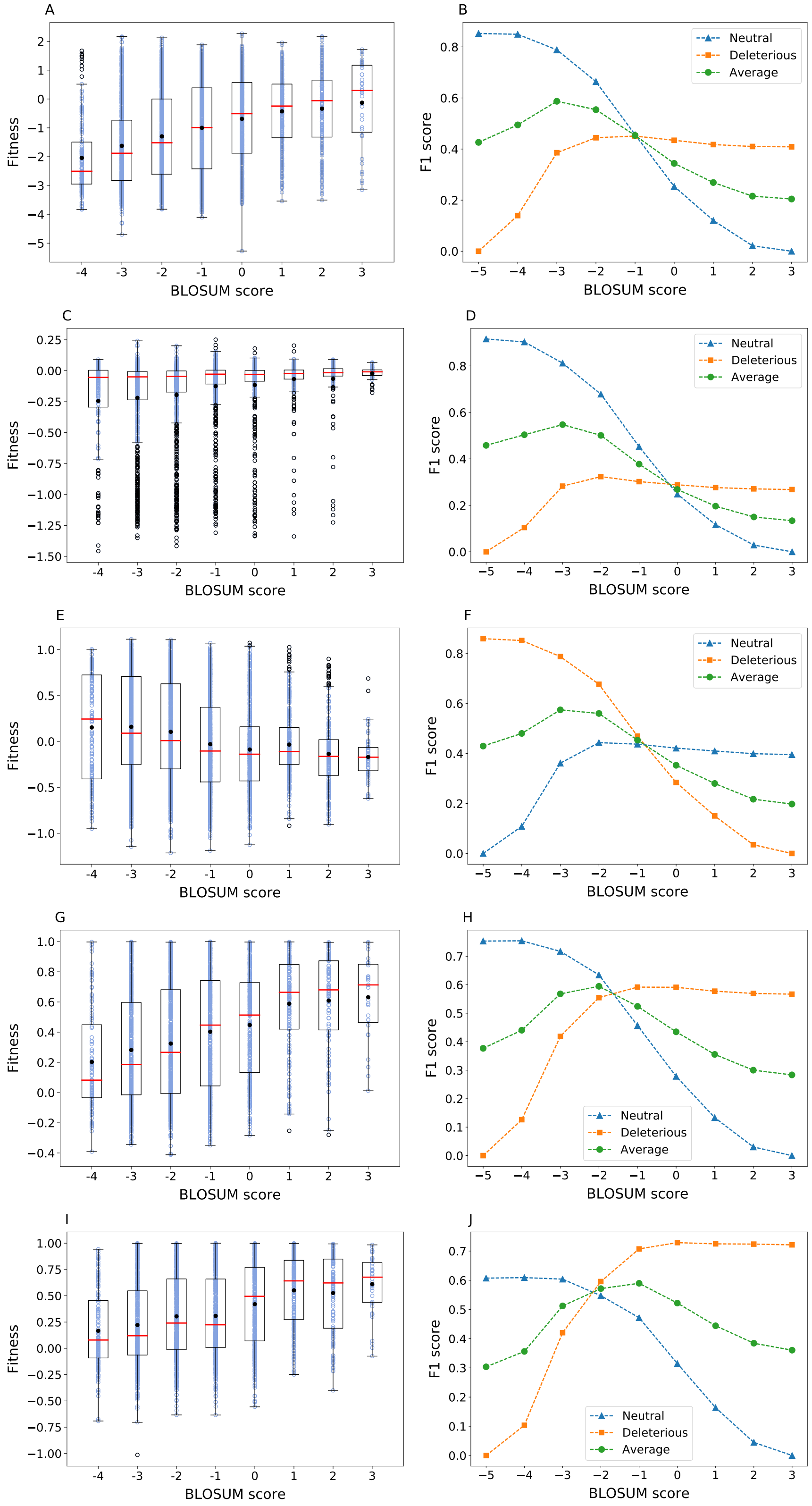

Figure E.8: Correlation between fitness and BLOSUM substitution matrix score and the F1 scores when the BLOSUM threshold for classifying mutations to neutral and deleterious is varied for APH(3')-II **(A,B)**, Hsp90 **(C,D)**, MAPK1 **(E,F)**, UBE2I **(G,H)** and TPK1 **(I,J)**. See Figure E.3 for details of box plot representation.

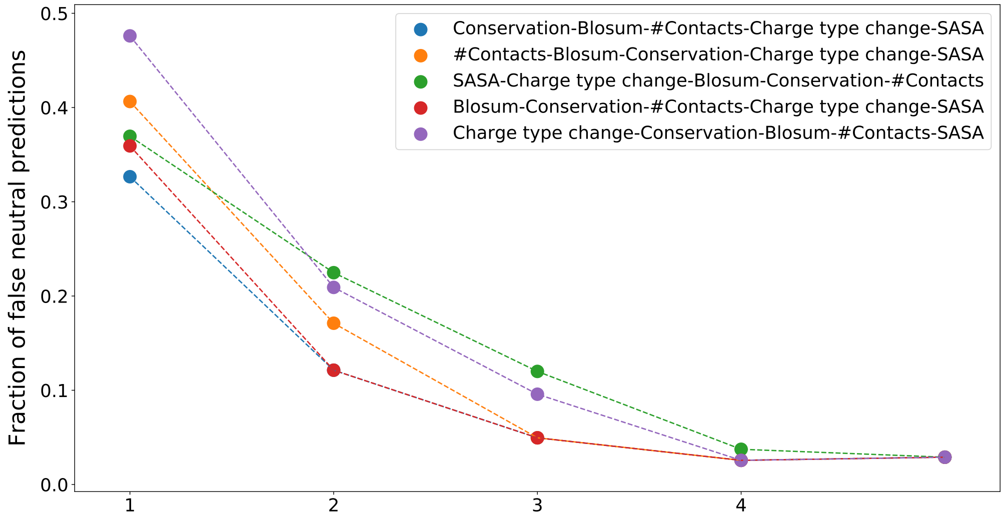

Figure E.9: Reduction in the number of mutations predicted wrongly as neutral as the number of criteria used is increased is shown for the case of β-lactamase. The order in which thresholds related to different variables are included is shown in the legend.

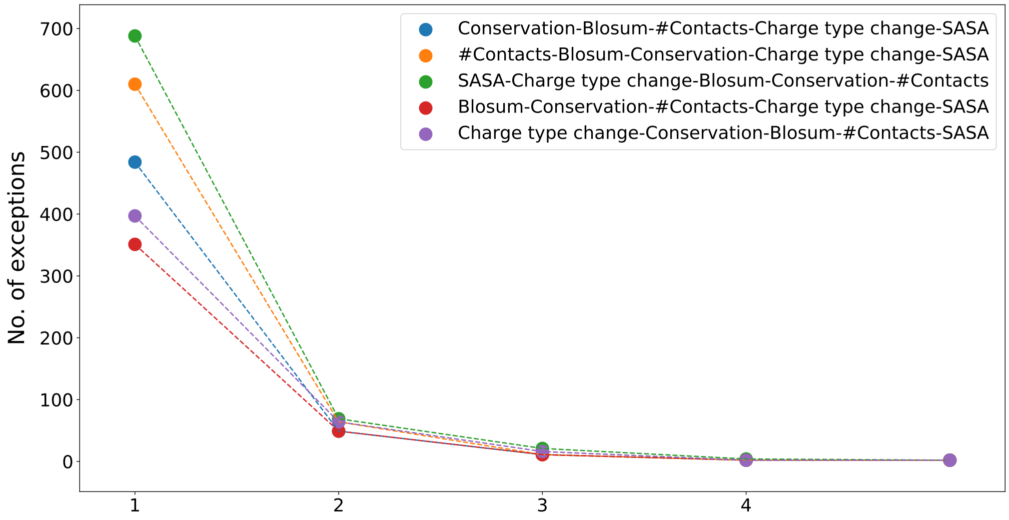

Figure E.10: Reduction in the number of mutations predicted incorrectly as neutral as the number of variables used is increased. Although the error fraction decreases as shown in Figure E.9, the number of mutations which are classified as neutral or deleterious by all variables also decreases.

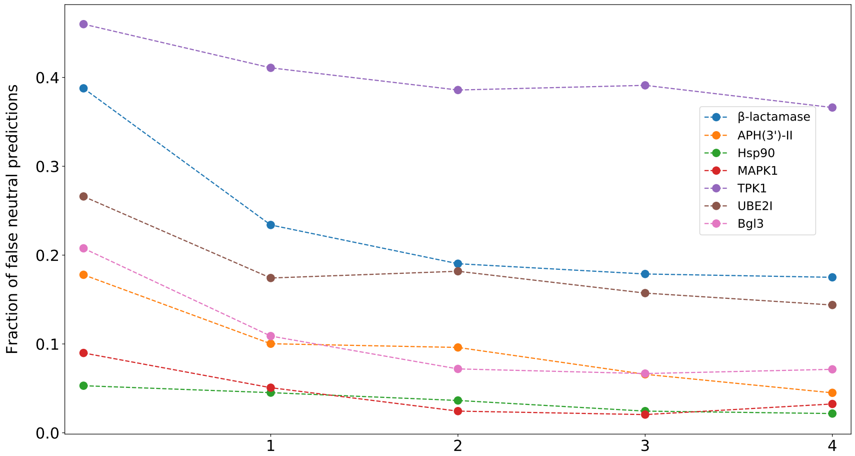

Figure E.11: Reduction in the chance of false-neutral predictions as the number of variables used for classification are increased. Threshold used for each variable is the average values given in **Table 1**. The trajectory for which the sum of error fractions is the least is shown for each protein. The order in which different threshold criteria are included for each protein is: 1) β-lactamase: SASA-Charge type change-Blosum-No.of contacts-Conservation, 2) APH(3')-II: Blosum-No. of contacts-Charge type change-SASA-Conservation, 3) Hsp90: Conservation-Charge type change-SASA-Blosum-No. of contacts, 4) MAPK1: Conservation-Charge type change-SASA-Blosum-No. of contacts, 5) TPK1: Charge type change-SASA-Conservation-No. of contacts-Blosum, 6) UBE2I: SASA-Charge type change-No. of Contacts-Blosum-Conservation, 7) Bgl3: SASA-Charge type change-No. of contacts-Blosum-Conservation.

Variable	Protein	True neutral	False neutral	True deleterious	False deleterious
Conservation	Beta-lactamase	1537	1731	625	59
	APH(3')-II	2621	649	439	525
	Hsp90	2306	129	493	1093
	MAPK1	2168	214	888	1200
	UBE2I	1357	636	378	192
	TPK1	909	1008	786	478
No. of contacts	Beta-lactamase	1071	962	1394	525
	APH(3')-II	2058	422	666	1088
	Hsp90	2235	364	258	1164
	MAPK1	1973	498	604	1395
	UBE2I	974	470	544	575
	TPK1	872	1047	747	515
SASA	Beta-lactamase	1268	803	1553	328
	APH(3')-II	2277	466	622	869
	Hsp90	2349	250	372	1050
	MAPK1	2041	326	776	1327
	UBE2I	1236	448	566	313
	TPK1	962	1085	709	425
BLOSUM	Beta-lactamase	1013	963	1393	583
	APH(3')-II	1752	379	709	1394
	Hsp90	1850	203	419	1549
	MAPK1	1915	374	728	1453
	UBE2I	893	373	641	656
	TPK1	821	793	1001	566

Table E.1: Number of true and false predictions for each protein using the average thresholds.

Variable	Protein	Threshold	True neutral	False neutral	True deleterious	False deleterious
Conservation	Beta-lactamase	0.65	1579	1917	439	17
	APH(3')-II	0.6	2621	649	439	525
	Hsp90	0.55	2167	102	520	1232
	MAPK1	0.5	1720	166	936	1648
	UBE2I	0.6	1357	636	378	192
	TPK1	0.6	909	1008	786	478
No. of contacts	Beta-lactamase	19	1083	1007	1349	513
	APH(3')-II	18	2058	422	666	1088
	Hsp90	17	2072	320	302	1327
	MAPK1	18	1973	498	604	1395
	UBE2I	19	1020	522	492	529
	TPK1	18	872	1047	747	515
SASA	Beta-lactamase	0.2	1268	803	1553	328
	APH(3')-II	0.2	2277	466	622	869
	Hsp90	0.2	2349	250	372	1050
	MAPK1	0.3	1809	281	821	1559
	UBE2I	0.2	1236	448	566	313
	TPK1	0.2	962	1085	709	425
BLOSUM	Beta-lactamase	-2	1013	963	1393	583
	APH(3')-II	-2	1752	379	709	1394
	Hsp90	-2	1850	203	419	1549
	MAPK1	-2	1915	374	728	1453
	UBE2I	-2	893	373	641	656
	TPK1	-2	821	793	1001	566

Table E.2: Number of true and false predictions for each protein when the threshold obtained by averaging the thresholds of other proteins (Leave One Protein Out analysis) is used. The threshold obtained by averaging over other 5 proteins is given in 3^{rd} column.

Protein	Variable	Predicted as neutral						Predicted as deleterious					
		th_{av}		th_{av}-δth		th_{av}+δth		th_{av}		th_{av}-δth		th_{av}+δth	
		TNP*	FFP*	TNP	FFP	TNP	FFP	TNP	FFP	TNP	FFP	TNP	FFP
β-lactamase	Conservation	200	0.18	180	0.16	207	0.18	203	0	309	0.04	124	0
	No of contacts			120	0.16	243	0.18			243	0.00	117	0
	SASA			242	0.21	152	0.13			203	0.00	203	0
	BLOSUM			125	0.08	220	0.19			267	0.00	100	0
Hsp90	Conservation	231	0.02	240	0.00	269	0.04	260	0.58	290	0.61	182	0.51
	No of contacts			160	0.03	272	0.03			357	0.59	107	0.38
	SASA			270	0.03	186	0.03			122	0.75	275	0.57
	BLOSUM			148	0.02	251	0.03			335	0.59	151	0.60
APH3'-II	Conservation	289	0.04	276	0.04	307	0.07	186	0.45	258	0.50	126	0.41
	No of contacts			206	0.03	344	0.07			236	0.47	107	0.50
	SASA			342	0.06	263	0.03			131	0.48	232	0.47
	BLOSUM			172	0.02	330	0.05			240	0.50	96	0.42
MAPK1	Conservation	447	0.40	485	0.42	380	0.37	185	0.03	152	0.03	215	0.03
	No of contacts			586	0.41	234	0.43			124	0.05	232	0.03
	SASA			203	0.33	514	0.43			257	0.03	157	0.04
	BLOSUM			571	0.44	256	0.38			122	0.02	199	0.04
UBE2I	Conservation	146	0.14	131	0.13	148	0.14	161	0.11	192	0.16	124	0.10
	No of contacts			90	0.16	195	0.14			184	0.13	103	0.11
	SASA			165	0.15	129	0.14			79	0.10	163	0.11
	BLOSUM			96	0.09	159	0.16			182	0.14	95	0.09
TPK1	Conservation	183	0.37	138	0.37	194	0.36	207	0.31	238	0.32	156	0.30
	No of contacts			134	0.36	207	0.37			280	0.29	87	0.29
	SASA			222	0.36	141	0.38			119	0.36	264	0.31
	BLOSUM			134	0.28	195	0.38			263	0.31	111	0.29

Table E.3: Table showing the sensitivity of quality of predictions with small variation in the thresholds. The change in threshold, δth was chosen to be approximately 10% of the maximum value of the given variable. When a variable threshold, th_{av} is varied by $\pm\delta th$, keeping all other variable thresholds same, the number of variants predicted as neutral/deleterious and the fraction of false prediction in each case are given. The column th_{av} is for the case when the average thresholds are used for all variables.
*TNP and FFP stands for total number of predictions and fraction of false predictions respectively.

Bibliography

[1] M. A. Stiffler, D. R. Hekstra, and R. Ranganathan, “Evolvability as a function of purifying selection in tem-1 β-lactamase,” *Cell*, vol. 160, no. 5, pp. 882–892, 2015.

[2] E. Firnberg, J. W. Labonte, J. J. Gray, and M. Ostermeier, “A comprehensive, high-resolution map of a gene’s fitness landscape,” *Molecular Biology and Evolution*, vol. 31, no. 6, pp. 1581–1592, 2014.

Appendix F

Interpreting Mutational Effects Predictions, One Substitution at a Time

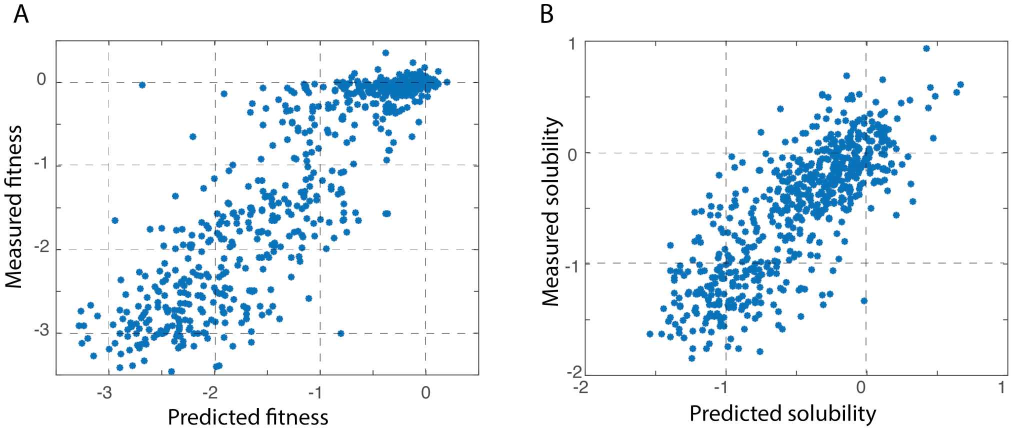

Figure F.1: Quality of predictions. 75% of the mutational data was used for training the models. The quality of the model was judged by comparing the predictions to the observations. The scatter plots show that both A. fitness and B. solubility can be predicted with a good reliability.

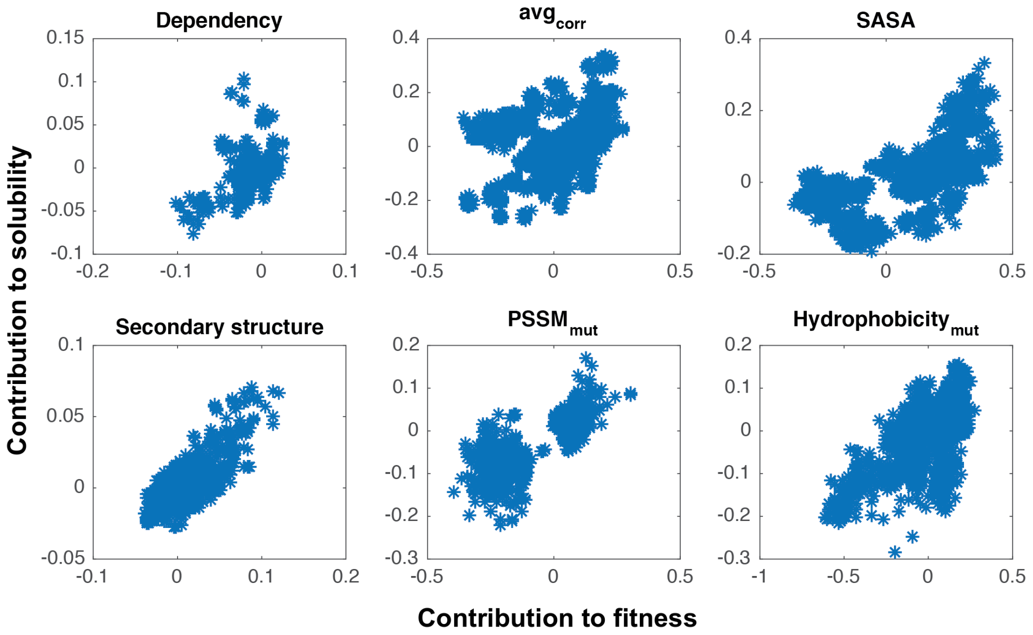

Figure F.2: Correlation of contributions to solubility and fitness. The SHAP values defining the contribution of each variable to fitness and solubility are shown. From the data it is clear that the contributions from these descriptive parameters to fitness (x-axis) and solubility (y-axis) are poorly correlated i.e., knowing a parameter contributes to an increase in fitness does not immediately clarify its possible contribution to solubility.

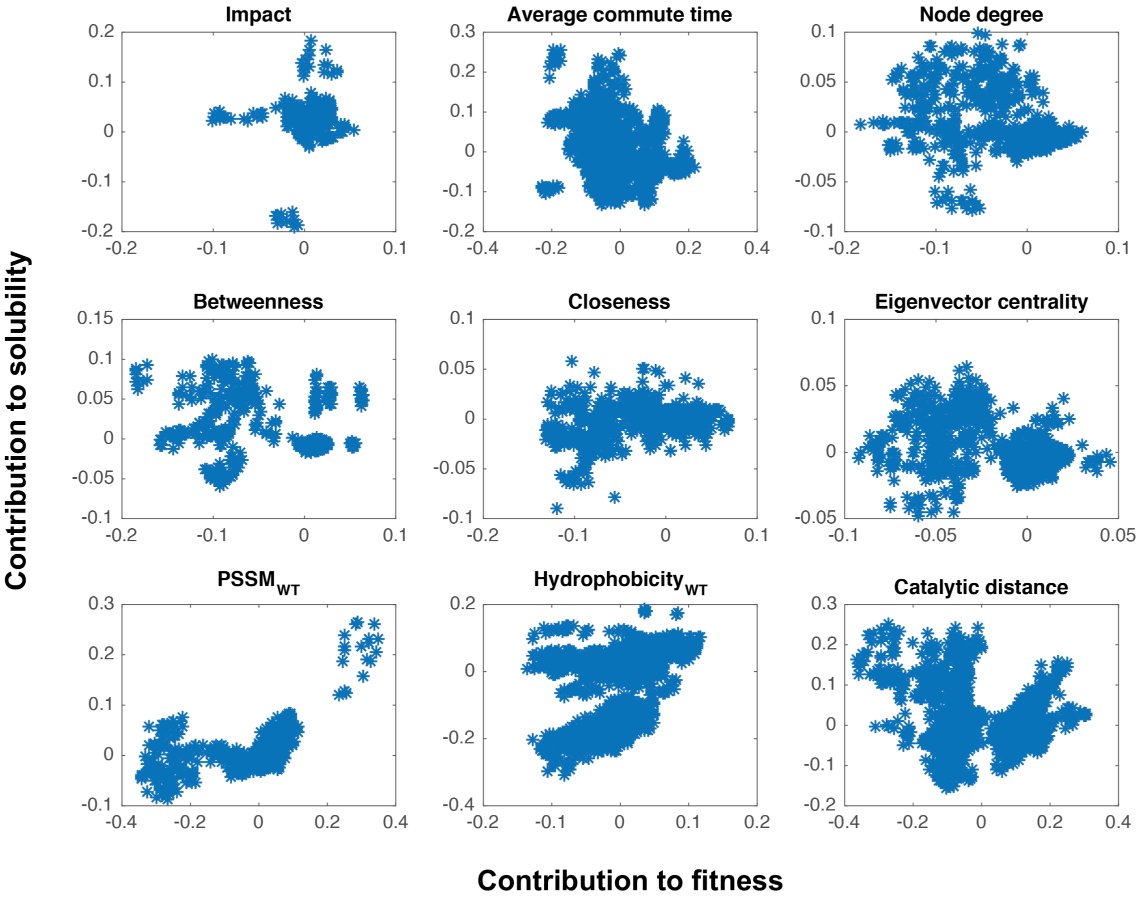

Figure F.3: Correlation of contributions to solubility and fitness. A continuation F.2, showing the contributions from some more parameters to fitness and solubility are poorly correlated.

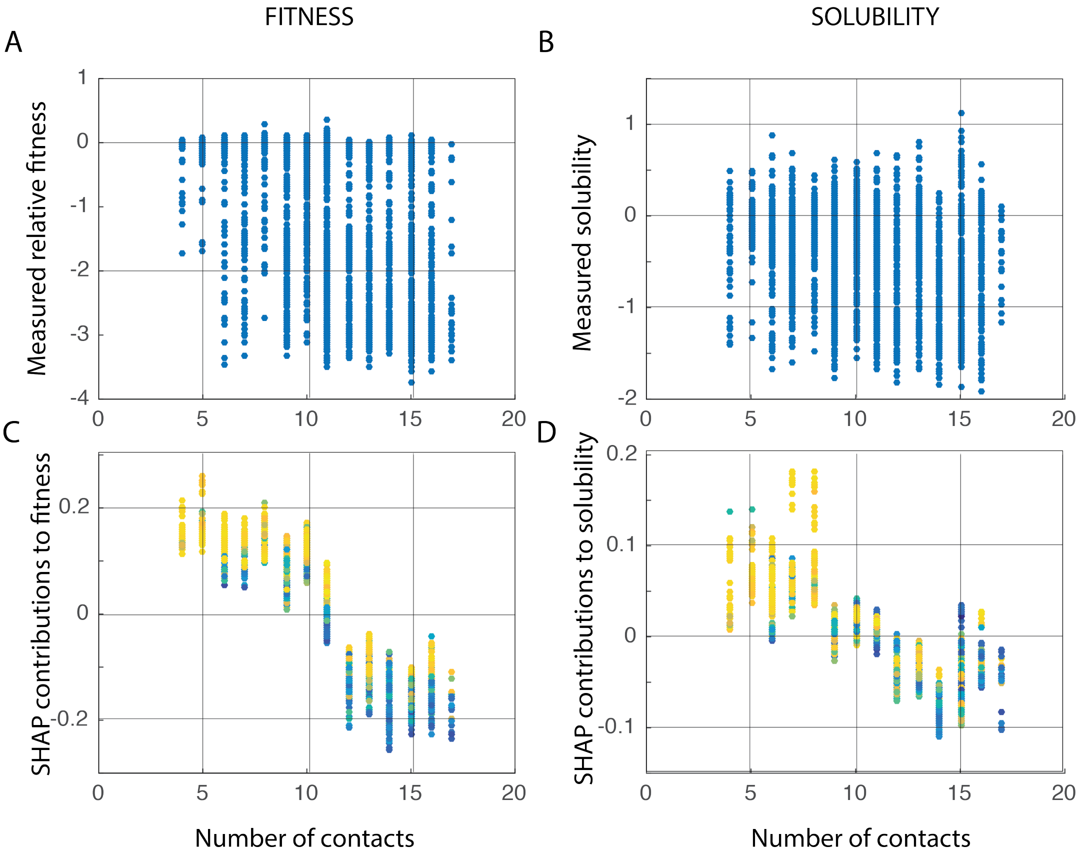

Figure F.4: Extracting relations. The scatter plots of A. fitness and B. solubility relative to the number of contacts the wild type amino acid has in its structure do not show a clear pattern. On the contrary, the SHAP contributions to C. fitness and D. solubility show a very clear pattern of reducing SHAP values with increasing contacts, which suggests that the fitness and solubility decrease with the substitution of a tightly packed amino acid. The colorbar is the same as in Figure 8.3 of Chapter 8, and represents the observed fitness changes.

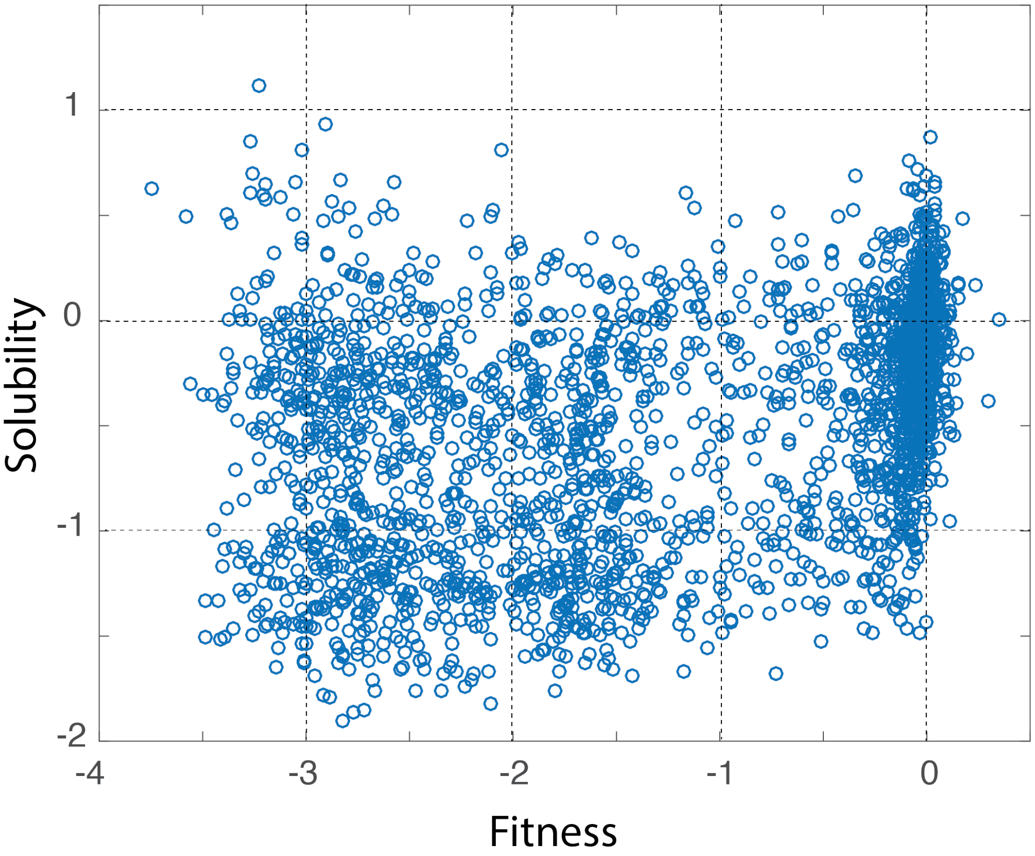

Figure F.5: Correlation between outcomes of multi-factorial contributions. The fitness and solubility from the experimental data is shown. While some of the factors contributed similarly to both solubility and fitness, some others in an opposite way, and many others in an uncorrelated way. Hence, the net result of the multi-factorial effect is a poor correlation between the fitness and solubility. However, seeing the nice patterns such as those in Figures 4, 5 it is apparent that the fitness and solubility can be reconstructed from the knowledge of the individual factors.

Appendix G

Using Deep Mutational Scan Data for Understanding Site Specific Codon Usage Bias

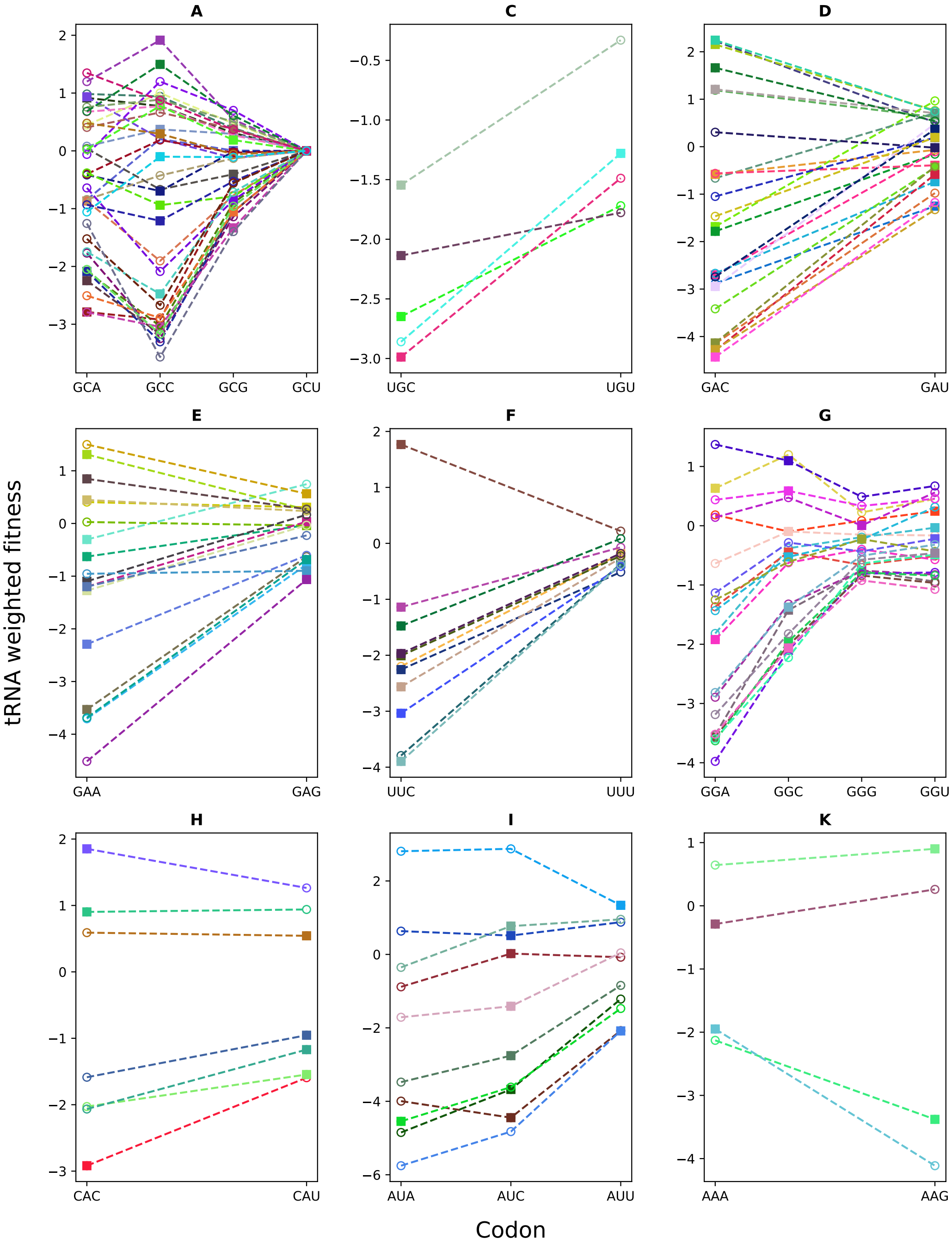

Figure G.1: The average of tRNA weighted fitness scores of variants that are possible with one nucleotide change for every synonymous codon of all amino acids except M and W in APH(3′)-II. Colours within each subplot correspond to different residue positions in the protein with the specific amino acid. Square symbol indicates the codon in the WT sequence. The results for the remaining amino acids are given in Figure G.2.

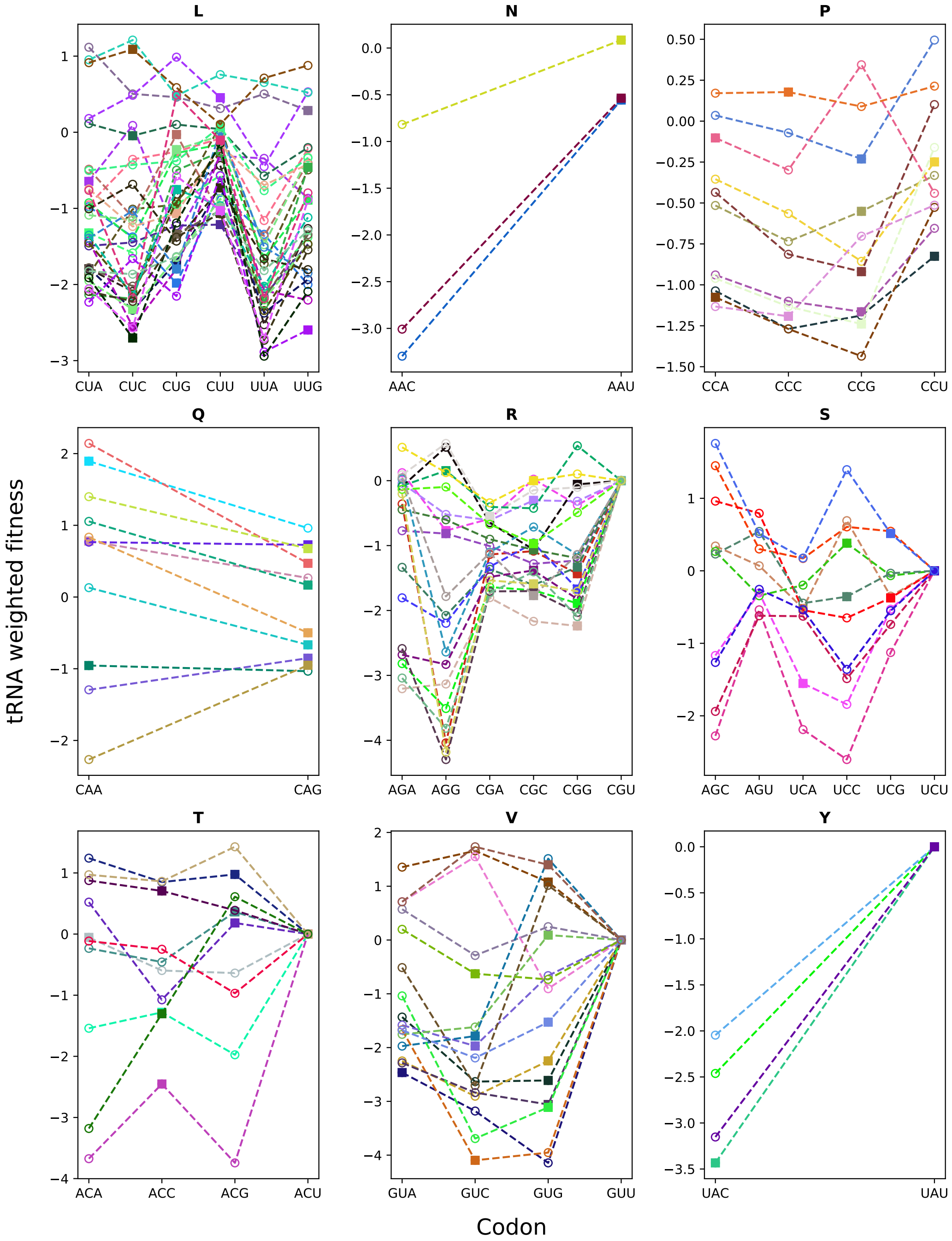

Figure G.2: See caption of Figure G.1.

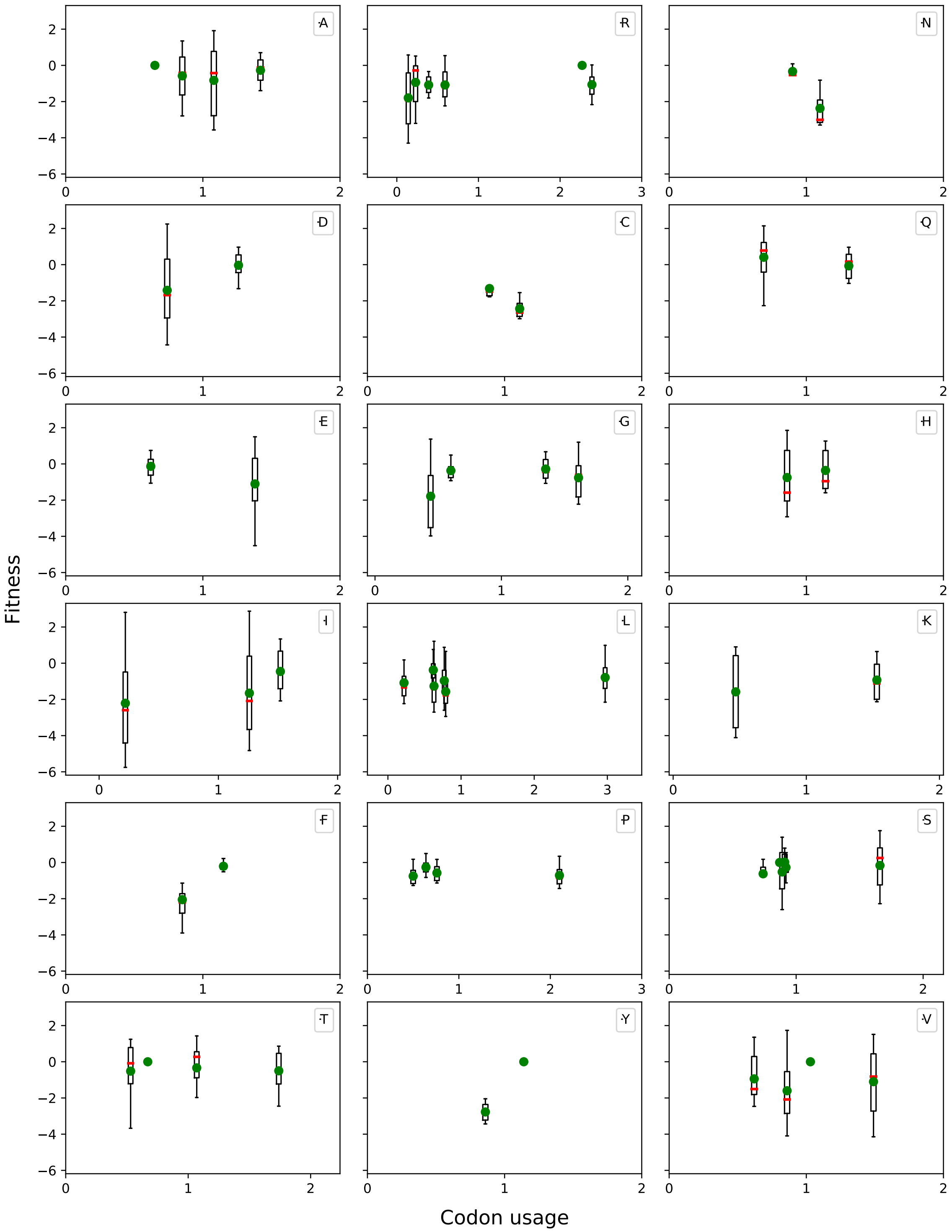

Figure G.3: The tRNA weighted fitness distributions plotted with respect to codon bias for the possible codons of each amino acid separately for APH(3′)-II.

		No. of amino acids	tRNA weighted fitness		tRNA availability	
			No. of correct predictions	Fraction of correct predictions	No. of correct predictions	Fraction of correct predictions
Secondary Structure	helix	113	36	0.32	42	0.37
	sheets	44	15	0.34	20	0.45
	coil	93	38	0.41	34	0.37
Chemical nature of amino acid	charged	71	34	0.48	31	0.44
	polar	53	15	0.28	19	0.36
	hydrophobic	126	40	0.32	46	0.37
Conservation	<= 0.25	31	12	0.39	8	0.26
	> 0.25	219	77	0.35	88	0.40
	<= 0.5	122	44	0.36	43	0.35
	> 0.5	128	45	0.35	53	0.41
	<= 0.75	181	66	0.36	68	0.38
	> 0.75	69	23	0.33	28	0.41

Table G.1: The fraction of correct predictions using each of the scoring method shown separately for amino acids in β-lactamase in different structural context, or having chemical different nature or functional importance as quantified using conservation.

		No. of amino acids	tRNA weighted fitness		tRNA availability	
			No. of correct predictions	Fraction of correct predictions	No. of correct predictions	Fraction of correct predictions
Secondary Structure	helix	96	26	0.27	35	0.36
	sheets	45	14	0.31	15	0.33
	coil	112	43	0.38	47	0.42
Chemical nature of amino acid	charged	74	28	0.38	30	0.41
	polar	39	15	0.38	13	0.33
	hydrophobic	140	40	0.29	54	0.39
Conservation	<= 0.25	81	32	0.40	28	0.35
	> 0.25	172	51	0.30	69	0.40
	<= 0.5	173	60	0.35	70	0.40
	> 0.5	80	23	0.29	27	0.34
	<= 0.75	219	72	0.33	82	0.37
	> 0.75	34	11	0.32	15	0.44

Table G.2: The fraction of correct predictions using each of the scoring method shown separately for amino acids in APH(3′)-II in different structural context, or having different chemical nature or functional importance as quantified using conservation.