Learning from Large-scale Mutagenesis Data

Vanessa E. Gray

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2017

Reading Committee:
Douglas M. Fowler, Chair
Richard Gardner
Christine Queitsch

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Learning from Large-scale Mutagenesis Data

Vanessa E. Gray

Chair of the Supervisory Committee:
Dr. Douglas M. Fowler
Genome Sciences

Mutations can have profound effects on protein function. For example, mutations can increase or decrease enzymatic activity, influence aggregation propensity, or lead to novel protein functions. Mastery of the rules governing how mutations affect protein function has the potential to revolutionize bioengineering. Recently, advances in DNA sequencing technologies and molecular biology techniques have afforded new methods, such as deep mutational scanning, to measure the quantitative effects of mutations on protein function in high throughput. In this dissertation, I first describe the state of the deep mutational scanning field. In the following chapters, I employ large-scale mutagenesis data sets from deep mutational scanning experiments to perform three investigations. In Chapter 2, I report how different amino acids affect the severity of mutational effect. In Chapter 3, I show how machine-learning algorithms can be used to model the evolutionary, structural and physicochemical properties of mutations from large-

scale mutagenesis data. In Chapter 4, I describe initial work on a deep mutational scan of amyloid β to reveal how mutations can affect the aggregation propensity of a protein. In Chapter 5, I discuss some outstanding questions that can be resolved with future analyses of large-scale mutagenesis datasets.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I owe great thanks to many. Firstly, I'd like to thank the department of Genome Sciences for its encouragement and support.

I thank Doug Fowler. I could not have found a more thoughtful advisor for my studies. I am tremendously thankful for his patience and guidance. Doug has a built a lab of not only talented scientists, but also kindhearted comedians who I looked forward to seeing everyday. I thank Ethan, who despite sitting next to me for ~2,000 days remained a cherished friend.

I thank Sudhir Kumar for his mentorship that inspired me to pursue a career in science.

I thank Sahana and Sanjay Srivatsan for showing me true friendship and Nick Hasle for immeasurable encouragement, love and adventure. Jeg elsker dig.

Also, I thank Suzanne Howard (my mother), Gilbert Gray (my father), and Paula Gray (my grandmother) for their support.

And finally, I thank the Seattle Bouldering Project because without it, I would not have sanity, upper body strength or befriended so many adventurous and down-to-earth people.

Chapter 1. INTRODUCTION

The ability of next-generation DNA sequencing to identify genetic variation in humans vastly outpaces geneticists' ability to interpret variation. For instance, a recent survey of genetic variation in humans revealed that each individual harbors ~50 missense mutations, many of which have unknown effects on health. To keep pace with the deluge of human variation, experimental methods have evolved to characterize the effects of mutations *en masse*. Such approaches offer high-resolution and quantitative information on the effects of mutations on protein function. Consequently, these new mutational datasets offer a hitherto untapped training source for models that predict the effects of mutations. To continue to improve our understanding of mutations and their effects, new assays must be developed to measure under-studied phenotypes, such as protein aggregation.

1.1 MUTAGENESIS

Making and studying mutants is a fundamental way to learn about proteins, revealing functionally important positions, validating specific hypotheses about catalytic mechanism and yielding insights into protein folding and stability. Site-directed mutagenesis is a widely used molecular biology technique to introduce codon mutations in deoxyribonucleic acid (DNA) sequence (1). In this Nobel Prize-winning technique, mutant DNA oligomers anneal to complementary template DNA and act as primers for polymerases to perform elongation (2). The resulting mutated DNA can be introduced into a model organism to identify phenotypic effects of mutated DNA or protein. Advanced approaches applied site-directed mutagenesis to stretches of DNA sequence to survey the effects of a mutation across a region of protein sequence. This

technique is called single amino acid scanning mutagenesis and has largely featured alanine (3). While site-directed and scanning mutagenesis have greatly improved understanding of molecular mechanisms in biology and disease, their throughput vastly underserves the need of contemporary geneticists.

1.2 DEEP MUTATIONAL SCANNING

Deep mutational scanning offers a high-throughput method for characterizing mutational effects (4, 5). Deep mutational scanning couples a functional assay with next-generation DNA sequencing to assess the effects of tens of thousands of mutations in a single experiment. While deep mutational scans can be adapted to measure different features of protein function, all deep mutational scans follow the same general workflow. First, a library of variants is created for a genetic region of interest. Libraries can be randomly or methodically designed to contain single, double and/or higher order amino acid mutations depending on the scientific questions that a deep mutational scan is intended to address (See Section 1.1.1). Second, a functional assay is used to increase the frequency of variants with a particular phenotype. Deep mutational scanning functional assays can take various forms and are described more deeply in Section 1.1.2. Third, next-generation sequencing is used to tabulate the changes in variant frequency before and after a selection assay is applied to score the effect of mutation(s) on protein function.

1.2.1 *Creating variant libraries*

Deep mutational scans require a library of variants. Variant libraries can be generated a number of ways, each technique with unique pros and cons (**Figure 1-1A**). The original deep mutational

scan used doped-oligomers to randomly[1] mutagenize the WW domain of human Yap65 (6, 7). Doped-oligo mutagenesis yields libraries with variable numbers of mutations per variant and it is nearly impossible to attain all possible single amino acid mutations of a mutagenized region due to bias against multi-nucleotide variants. Another method, called site-saturation mutagenesis, most effectively generates variant comprehensive single amino acid variant libraries (8). This technique requires a unique polymerase chain reaction (PCR) for each position within a mutagenized region and thus can be very time consuming.

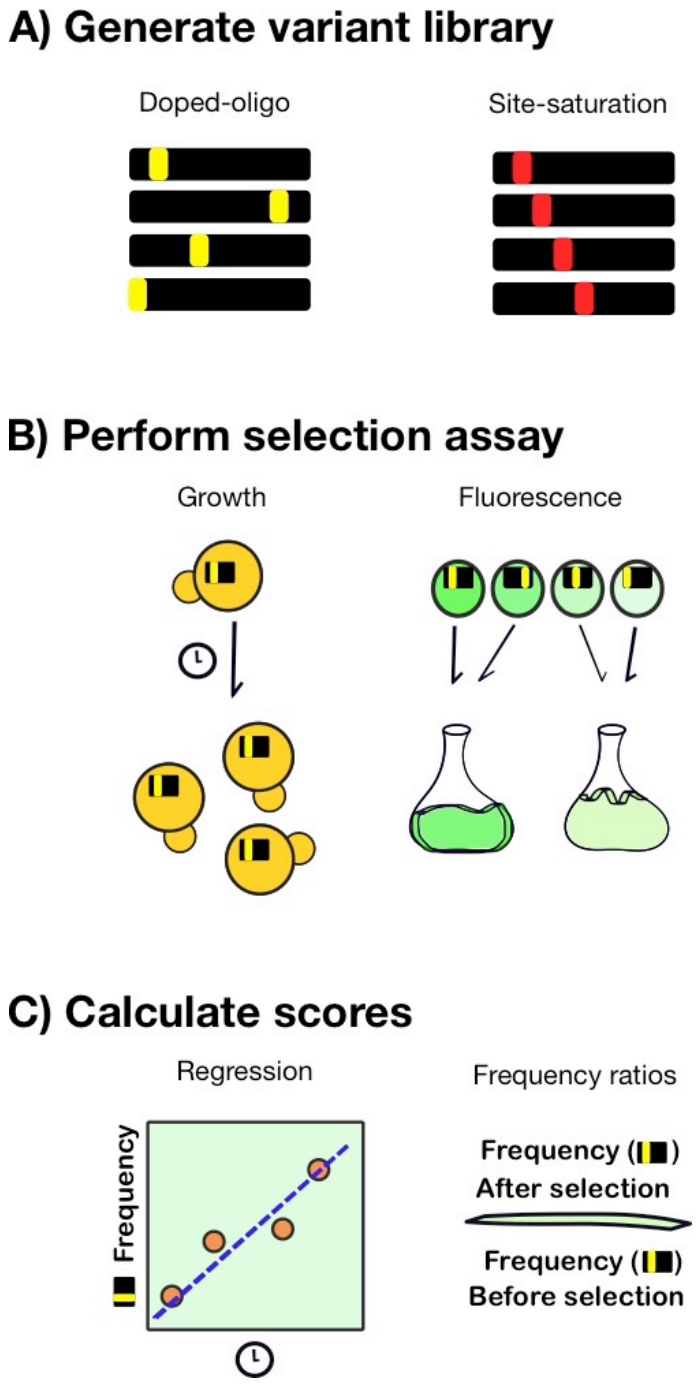


Figure 1-1 Overview of deep mutational scanning workflow. (A) First, a DNA variant library is made. Previous studies have commonly used a doped-oligo or site-saturation mutagenesis protocols. (B) Next, a functional assay is used to enrich variants with a particular phenotype. (C) Variant effect scores can be calculated using the slope of a regression on variant frequency over time or frequency ratios. Frequency ratios divide the relative frequency of a mutation before and after selection or in one group versus another.

[1] Doped-oligo libraries are not random at the protein level. While the nucleotide mutations are distributed randomly across DNA sequences, doubly and triply mutated codons are not well represented in these libraries.

1.2.2 *Assays to select for protein function*

Since the discovery of site-directed mutagenesis in 1978, functional assays have been routinely used to investigate the effects of mutations on protein function (2). Functional assays can work at the protein, subcellular, cellular or organismal level and be designed to measure a vast range of phenotypes. Such assays can be used to quantitatively discriminate between mutations that enhance, disrupt or maintain protein function and are immediately compatible with a deep mutational scanning framework (**Figure 1-1B**).

Here, I discuss three published functional assays that range in scope from generalized to highly specific. First, in a deep mutational scan of *heat shock protein 90* (HSP90) in yeast, a very simple, but effective assay was used (9). HSP90 is essential for yeast growth and strongly linked to organism fitness. Thus, yeast with nonfunctional HSP90 will die and decrease in frequency over time. Moreover, yeast growth can serve a proxy for the activity level of each variant. Growth assays are commonly employed for essential genes (10-14). However, when a protein's function is not linked to growth, or when a specific protein function is the target of study, an assay can be designed to artificially link a variant's activity to cellular growth. Such an assay was used to measure substrate (BARD1- RING domain) binding by the human BRCA1 RING domain (15). In Starita *et al.*'s assay, a Gal4 transcription factor was fused to the BRCA1 domain and the Gal4 activation domain was fused to BARD1. Moreover, upon BRCA1-BARD1 binding, the transcription factor and activation domain drives expression of a selection marker, such that yeast expressing BRCA1 variants that bind to BARD1 increase in abundance during the selection, while nonfunctional variants decrease. Thus, even though BRCA1 binds >25 interaction partners (16), this assay selectively only measures the ability of this protein to solely

bind BARD1. Additionally, assays need not be growth based. Activity can be quantitated via a fluorescence-based assay. In an assay designed to study the effects of mutations on GFP fluorescence (17), fluorescence–activated cell sorting is used to bin variants with similar fluorescence intensities. In other assays, where a naturally fluorescing protein is not the subject if study, fluorescence can be linked to variant activity, such that high activity variants yield high fluorescence, and low words, activity variants yield low fluorescence (e.g. (18, 19).

1.2.3 *Next-generation sequencing and quantifying the effects of mutations on protein function*

High-throughput DNA sequencing is a cornerstone of deep mutational scanning (5, 20, 21). One type of commonly used sequencing is the Illumina platform (22, 23), which offers short-reads with relatively low error and cost. To prepare a variant library for Illumina sequencing, a region containing the sequence of interest is PCR amplified to form amplicons with small DNA adapters on the 3' and 5' ends. To sequence, a DNA sample is washed over a flow cell, which contains short DNA oligos. These DNA oligos anneal to homologous regions within the adapter region. Through a process called 'bridge amplification', the DNA is amplified to form 'polonies', otherwise known as clusters on the flow cell. Next, the sequencing process begins by washing the flow cell with reversible, fluorescence-labeled terminators, primers and DNA polymerase. To determine the first base, a camera captures the emitted fluorescence of a polony after excitation and calls the nucleotide associated with the associated wavelength. The camera can capture the emissions of tens to hundreds of millions polonies at once. Next, the terminator is cleaved, the next terminator is added and the imaging repeats until a programmed stopping point is reached. Upon completion, the image data is used to spell out the order of nucleotide bases in a DNA sequence.

Like all tools, Illumina sequencers have limitations. For instance, the Illumina NextSeq 550 can read a maximum of 300 base pairs per read with 1 erroneous base per 1,000 bases (23-25). Thus, sequencing through an average human gene, which contains ~900 nucleotides becomes impossible (26). Barcoding is a proven method that overcomes this limitation (27). In barcoding, a short randomized DNA sequence (usually < 20 bases) is cloned up- or down-stream of the mutagenized region. With barcodes, deep mutational scanning can be carried out as explained above, however in order to track variant frequencies before and after the selection assay is applied, barcodes are sequenced rather than variants. To create a map of barcode-variant pairs, a sequencing platform with longer read potential, such as an Illumina MiSeq (24) or PacBio (28) sequencer is used to read through the mutagenized region and it's cognate barcode. Thus, barcoding and subassembly can be used to overcome read length limitations of short-read sequencing platforms.

Several methods for scoring variant functions from variant frequencies exist. If variant frequency is only available at two timepoints, variants can be scored using a ratio of frequencies after and before selection (**Figure 1-1C**). When more than two timepoints exist, variants can be scored by regressing on mutant frequency versus time and calculating the slope. Computational tools, such as Enrich2 (29), calculate ratio and regression scores and offer a platform for standardized data analysis of deep mutational scanning sequencing data.

1.3 Variant Effect Prediction

Experiments can reveal a variant's molecular effect, and recent advances in multiplex assays have enabled the assessment of large numbers of variants (4, 30). However, we are far from having a comprehensive atlas of missense variant effects in the human proteome, and such an atlas is a distant goal for model organisms. To address this challenge, computational and evolutionary biologists have designed variant effect predictors to help interpret the effects of mutations. Early predictors were created to combat the inundating volume of uncharacterized genetic variants gathered by genome sequencing initiatives (31-33). Because diagnosing variant pathogenicity was a high priority, predictors were designed to categorize a mutation either as pathogenic/deleterious or benign.

While some predictors are simple statistical models (32, 34, 35), others are products of sophisticated supervised machine learning algorithms. Machine learning is the use of an algorithm to model relationships between a response variable and descriptive features. Response variables can either be discrete or continuous and are modeled by classification and regression models, respectively. For instance, the response variables used to train the vast majority of mutational predictors were discrete: disease- and nondisease-associated mutations (36-50). Descriptive features vary widely between predictors, but generally describe structural, evolutionary and/or physicochemical properties of amino acid positions. How the response variable and descriptive features are modeled is dependent upon the particular machine-learning algorithm used. For example, gradient boosting algorithms aim to train an ensemble of weighted decision trees. Each tree's training data is either randomly sampled or enriched for difficult-to-

predict data depending on whether stochastic gradient boosting (51) or the Adaboost (52) algorithm is used.

The response variable and descriptive features used to train models, yield models suited for a particular prediction task. For instance, the PolyPhen2 HumDiv model is a support vector machine trained on thousands of human Mendelian disease-associated and neutral variants, and is therefore optimized to predict the clinical variant effects (53). SNAP2, an ensemble of neural network models, is trained on human pathogenic and neutral variants as well as variants that impact molecular function (42). Given the breadth of training data, SNAP2 predictions encompass both the clinical and molecular effects of missense variants. Conversely, SIFT and Evolutionary Action are not products of machine learning, but instead rely on evolutionary patterns to predict variant effects. Despite their simplicity, SIFT and Evolutionary Action perform similarly to PolyPhen2 and SNAP2 (38) , which highlights the importance of evolutionary information to successful variant effect prediction. A recently described unsupervised method, EVmutation, leverages evolutionary signatures of epistasis to predict variant effects, and has demonstrated enhanced accuracy over SIFT and PolyPhen2 for both molecular and clinical effect prediction (49). These tools are all used to prioritize variants in clinical and laboratory settings. However, many tools have severe predictive limitations. For example, many predictors show poor prediction accuracy for neutral mutations, particularly at evolutionarily conserved protein positions (54), which stems from biased training data, as noted by Liu and Kumar (55).

1.4 Assays to Measure Protein Aggregation

Variant effect prediction can be used to gain insight into the general effects of mutations, however, when we want to understand specific phenomenon we need to empirically measure the effects of mutations. One such feature of proteins that is not well understood is protein aggregation.

Amyloid is a fibrous cross-β structure formed from protein aggregates (56-59) . Although functional amyloid has been reported [4], pathological forms are involved in neurodegenerative diseases such as Alzheimer's disease (AD) (60, 61). AD is the most common form of dementia and is estimated to afflict over 6 million Americans, the vast majority of which are over 65 years of age (62, 63). Several causative genes are known for AD, however, only 1-10% of AD cases are due to dominantly inherited forms (62). Currently, no available treatment options exist, despite large-scale pharmaceutical endeavors to halt disease progression and/or repair diseased brains (64). This fact is particularly troubling considering that the number of Alzheimer's patients is projected to nearly triple by 2050.

The amyloid-forming peptide in AD is amyloid β (Aβ) (65, 66). This peptide is post-translationally cleaved from amyloid β precursor protein, which is a transmembrane glycoprotein commonly expressed in nearly all human cell types (67-69). This precursor protein is first cleaved by β-secretase to reveal the N-terminus of Aβ and then a γ-secretase complex releases the C-terminal end of Aβ (68, 70, 71). γ-secretase cuts the precursor protein at variable positions to produce peptides that range from 38 to 43 amino acids in length; the most aggregation-prone form of Aβ contains 42 amino acids ($A\beta_{42}$), though $A\beta_{40}$ is 10-fold more common than $A\beta_{42}$ (65,

72, 73). β-secretase and γ-secretase cleavage occurs as APP traverses the secretory pathway at subcellular membrane locations (*i.e*. endoplasmic reticulum, golgi, endosomes (70)) and thus Aβ can begin aggregating intraneuronally, though AD amyloid plaques are routinely found in the extracellular milieu (66, 74-77).

The aggregation of Aβ begins with a shift in equilibrium from soluble monomers to oligomeric peptides/proteins and these oligomers may serve as nucleation seeds for amylogenic fibrils. In AD, fibrils accumulate in the extracellular space to form amyloid plaques, which is a defining feature of the disease. Evidence from *in vitro* (78), cell culture (79), animal [40, 43-45] and postmortem brain (78, 80) studies support the hypothesis that oligomeric forms of Aβ are more neurotoxic than fibrils and plaques. Due to the role of Aβ aggregates in disease pathology, many studies have sought to characterize oligomeric and fibril structures.

Models of Aβ structure propose two β-strands that form parallel, in-register β-sheets (81). Due to the noncrystalline nature of amyloid fibrils, traditional techniques such as X-ray crystallography and solution-state NMR cannot be used. Instead, structural models are developed by amassing constraints, which has historically included: parallel, in register β-sheets from solid-state nuclear magnetic resonance, hydrogen-bonding constrains from hydrogen/deuterium exchange nuclear magnetic resonance and side-chain packing data from pairwise mutagenesis experiments (82). Many models are also underpowered because they are almost exclusively generated from *in vitro*-derived constraints for model building (83), which may not be representative of *in vivo* conditions.

Several years ago, *Saccharomyces cerevisiae* was successfully developed as a model organism for AD(84-86). Studies in this organism have informed follow-up studies on protein localization, interaction partners, nucleation process, and risk factors for disease. Recently, Morell *et al.* (86) developed a yeast-based system to extricate toxicity from aggregation and thus offers a way to investigate the effects of protein sequence on aggregation propensity alone. In this system, Aβ is cytoplasmically localized to eliminate its aggregation-associated toxicity (86). To link Aβ aggregation to yeast growth, Aβ is fused to an essential protein, dihydrofolate reductase (DHFR) via a short peptide linker. The endogenously expressed concentration of DHF1 (yeast DHFR), is competitively inhibited by methotrexate. Furthermore, DHFR's activity is dependent upon Aβ solubility, and thus yeast with soluble Aβ variants rapidly grow in culture, whereas aggregating Aβ variants yield slow yeast growth. As described in chaper 4, I leveraged this assay's ability to measure aggregation without potential confounding toxicity phenotypes. This assay is applied within a deep mutational scanning framework to study all single and some double amino acid mutations of Aβ.

Chapter 2. ANALYSIS OF LARGE-SCALE MUTAGENESIS DATA TO ASSESS THE IMPACT OF SINGLE AMINO ACID SUBSTITUTIONS

2.1 INTRODUCTION

Making and studying mutants is a fundamental way to learn about proteins, revealing functionally important positions, validating specific hypotheses about catalytic mechanism and yielding insights into protein folding and stability. Single amino acid scanning mutagenesis, in which every position in a protein is sequentially mutated to one particular amino acid, was a key advance. By searching sequence space systematically, scanning mutagenesis enabled the unbiased identification of positions and amino acid side chains important for protein function. The first application of scanning mutagenesis used alanine substitutions to identify positions in human growth hormone important for receptor binding(3). Alanine was chosen because it represents a deletion of the side chain at the β-carbon. In addition to alanine, many other amino acids including arginine (87), cysteine (88), glycine (89), methionine (90), phenylalanine (91), proline (92) and tryptophan (93) have been used for scanning mutagenesis, often with a specific hypothesis in mind (*e.g.* that bulky amino acids are important). Nevertheless, some suggest that alanine substitutions are especially useful for identifying functionally important positions or that they best represent the effects of other substitutions (94, 95).

Which amino acid best represents the effect of other substitutions? Which substitutions are ideal for finding functionally important positions, such as those that participate in binding interfaces? Answering these questions is important because single amino acid scanning mutagenesis continues to be used to understand and engineer proteins. Despite the large investment in scanning mutagenesis, little work has been done to systematically compare the effects of

different substitutions. Some scanning mutagenesis studies compare two different types of scans (*e.g.* alanine and cysteine), but generally find that the information revealed by each substitution is distinct (91, 96). Computational predictions for all substitutions at 1,073 positions across 48 proteins in the Alanine Scanning Energetics Database suggested that alanine substitutions correlated best with the mean effect of every mutation at each position (97). However, concrete answers to these questions require comparing the empirical effects of different substitutions in many proteins. Thus, I analyzed large-scale experimental mutagenesis data sets comprising 34,373 mutations in fourteen proteins. I found that proline is the most disruptive substitution

and methionine is the most tolerated. Global and position-centric analyses revealed that histidine and asparagine substitutions best represent the effects of other substitutions. I evaluated the utility of each amino acid substitution for determining whether a position is in a ligand-binding interface, and found that highly disruptive substitutions like aspartic acid, glutamic acid, asparagine and glutamine performed best. Thus, these results suggest that histidine and asparagine are the most representative substitutions, while aspartic acid and glutamic acid are the best choices for finding ligand-binding interfaces.

2.2 RESULTS

2.2.1 *Data collection and curation*

I curated sixteen large-scale mutagenesis data sets from published deep mutational scans of fourteen proteins (**Fig. 2.1A, Table 1)**. Here, I included two distinct data sets for the BRCA1 RING domain and for UBI4 because mutations in these proteins have been independently assayed for different protein functions (*e.g.* BRCA1 BARD1 binding and E3 ligase activity). This collection of data sets is ideal for an unbiased analysis of the general effects of mutations

because the mutagenized proteins are highly diverse, encompassing enzymes, structural proteins and chaperones from organisms ranging from bacteria to humans. The frequency of amino acids in the wild type sequences of the fourteen proteins was similar to amino acid frequencies in all known proteins(26) (**Fig. 2.1B**). For example, leucine (frequency = 11%) and alanine (8%) were

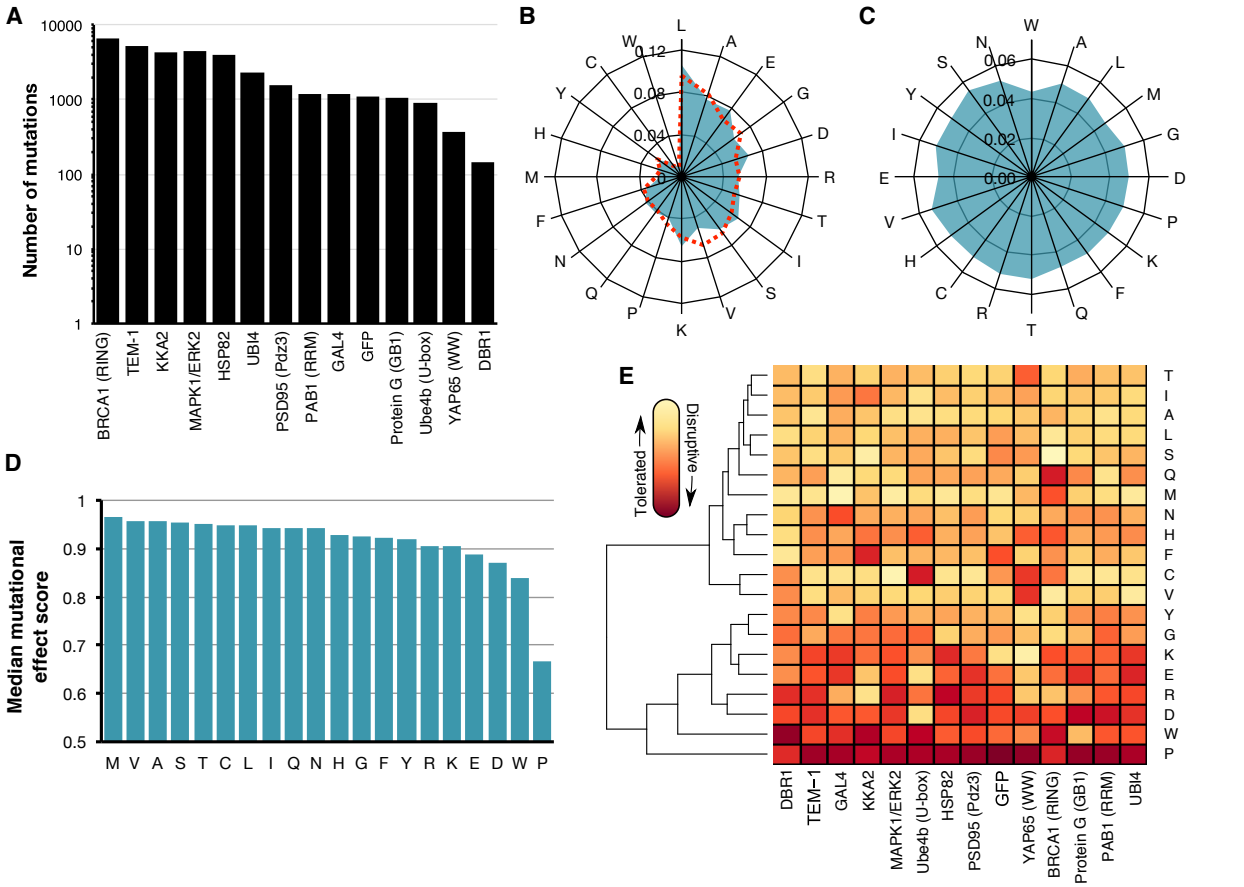


Figure 2-1. Large-scale mutagenesis data from fourteen proteins. (**A**) The number of single amino acid mutations with effect scores in each of the fourteen proteins is shown. (**B**) A radar plot shows the relative frequency of occurrence of each amino acid in the wild type sequences of the fourteen proteins (blue) or in 554,515 proteins in the UniProt Knowledgebase (26) (dashed red). (**C**) A radar plot shows the relative frequency of each of the twenty amino acid substitutions in the large-scale mutagenesis data sets for all fourteen proteins. (**D**) The median mutational effect score of each amino acid substitution is shown for 34,373 mutations at 2,236 positions in all fourteen proteins. (**E**) A heat map shows the median mutational effect score of each amino acid substitution for each protein separately. Yellow indicates tolerated substitutions while orange indicates disruptive substitutions. Amino acids and proteins were ordered according to similarity using hierarchical clustering with the hclust function from the heatmap2 package in R.

the most frequently occurring wild type amino acids in the fourteen proteins, while tryptophan (<1%) was the rarest. However, the unbiased and massively parallel nature of deep mutational scanning experiments yielded a relatively uniform distribution of amino acid substitutions (**Fig. 2.1C**). Furthermore, the data sets were generated by different labs at different times using different types of assays, reducing the chances of bias arising from specific experimental or analytical practices. Importantly, the assay formats used for the deep mutational scans included many commonly employed in alanine scanning like phage display and yeast two-hybrid. Collectively, these large-scale mutagenesis data sets comprised 34,373 nonsynonymous mutations at 2,236 positions in the fourteen proteins. The data sets contained effect scores for most mutations at each position. To facilitate comparisons between each data set, I rescaled mutational effect scores for each protein, using synonymous mutations to define wild type-like activity and the bottom 1% of mutations to define lack of activity (**Appendix A 1A; see Methods**). Thus, each mutational effect score reflects the impact of the mutation, relative to wild type, with a score of zero meaning no activity and a score of one meaning wild type-like activity.

2.2.2 *General patterns of mutational effect*

To validate the large-scale mutagenesis data, I examined expected patterns of mutational effect. For example, mutations to proline should generally disrupt protein function, as proline restricts the conformation of the polypeptide backbone and eliminates the amide hydrogen necessary for hydrogen bonding. Indeed, proline substitutions were overwhelmingly more disruptive than other substitutions to protein function (**Fig. 2.1D; Appendix A 1B**). In fact, proline was the most disruptive substitution in eleven of fourteen proteins and second most disruptive in the remaining three proteins (**Fig. 2.1E**). Additionally, as expected from the Dayhoff (98), Blosum (99) and

Grantham (100) substitution matrices, tryptophan tended to be deleterious. Methionine was the best-tolerated substitution and therefore may be useful for identifying the most immutable protein positions. Interestingly, mutations to alanine, which is commonly employed in scanning mutagenesis, were better tolerated than many other substitutions. Other substitutions were also well-tolerated, with seven different amino acids appearing as the most tolerated across the fourteen proteins (**Fig. 2.1D, E**). Tolerance to substitutions depends on structural context, so the variability in the best-tolerated substitution might be due to diversity in the structural composition of each protein in our data set. Thus, the large-scale mutagenesis data sets I collected generally recapitulated my expectations about the effects of mutations, despite coming from fourteen distinct proteins that were each assayed independently.

2.2.3 *Identifying representative amino acids from large-scale mutagenesis data sets*

Next, I determined which amino acid substitution best represented the effects of all other substitutions. To avoid bias arising from missing data, I restricted this analysis to the 882 positions in the fourteen proteins with measured effects for all nineteen possible substitutions. I calculated the median mutational effect at each of these 882 positions. Overall, the median effects across these positions were mildly disrupting, with a mean of 0.82 (stop ~ 0, wild-type ~ 1). I found that the effects of phenylalanine, glycine, histidine, isoleucine, leucine, asparagine, glutamine and tyrosine substitutions were all indistinguishable from the median effects (**Fig. 2.2A, Appendix A Table 1**). However, proline, aspartic acid and tryptophan substitutions were much more disrupting than the median substitution. Alanine, cysteine, methionine, serine, threonine and valine were considerably less disrupting than the median substitution. These well-tolerated amino acid substitutions might be useful for detecting the most mutationally sensitive

positions in a protein. However, these substitutions are not especially representative of the effects of other substitutions.

I also examined the dispersion of each amino acid's mutational effect about the median at all 882 positions, reasoning that representative substitutions would have minimal dispersion. Of substitutions whose effects were indistinguishable from the median effect, histidine and asparagine have the smallest dispersion (standard deviation = 0.15 and 0.14, respectively; **Fig. 2.2B**), while tyrosine (0.18), glutamine (0.16), phenylalanine (0.19), glycine (0.17), leucine (0.17) and isoleucine (0.19) all had larger dispersions. Thus, of all possible substitutions, histidine and asparagine tended to have effects closest to the median effect at the 882 positions I examined.

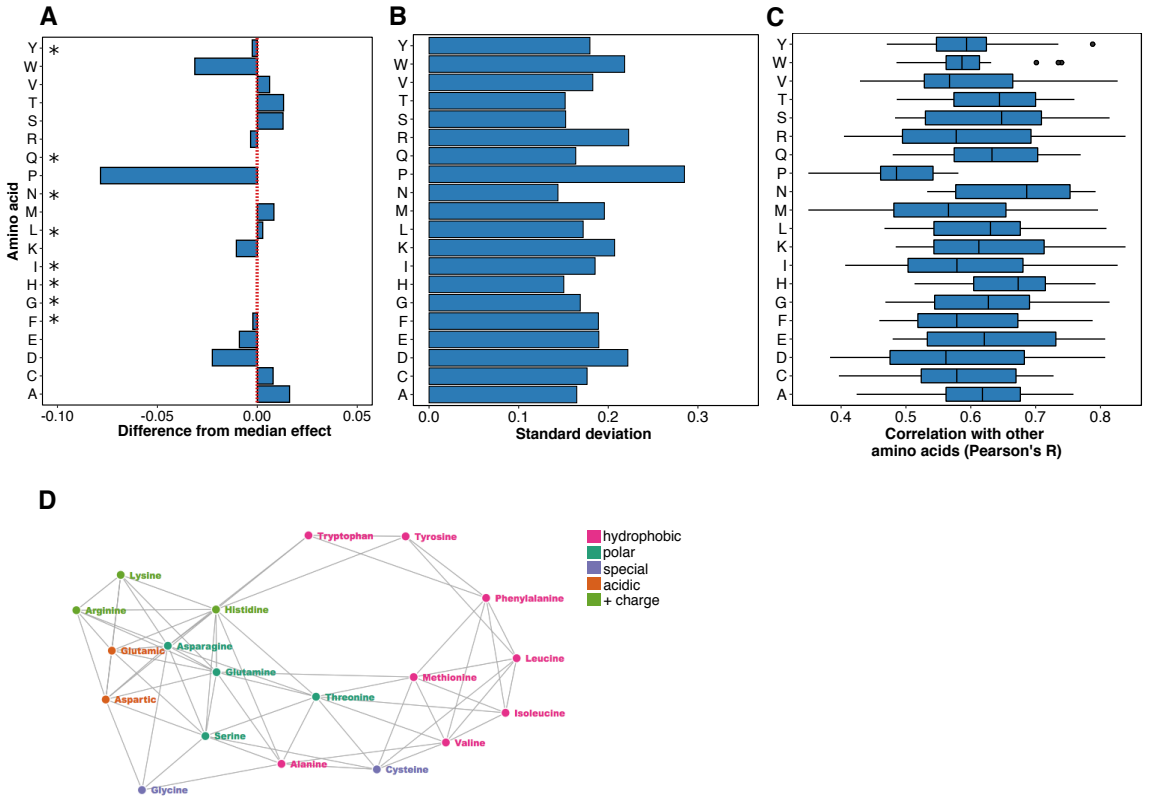


Figure 2.2. Histidine and asparagine substitutions best represent the effect of other substitutions. (A) For each of the 882 positions where the mutational effects of all nineteen substitutions were measured, the difference from the median effect was calculated for each substitution at each position. The median of these differences across all positions for each substitution is shown, with the red line indicating a median difference of zero. A paired, two-sided Wilcoxon rank sum test was used to determine whether each substitution's difference from the median effect across all positions was equal to zero (* indicates substitutions with a Bonferroni-corrected p-value > 0.01; Table S1). (B) The standard deviation of each substitution's differences from the median effect at the 882 positions where the mutational effects of all nineteen substitutions were measured is shown. (C) For each substitution, Pearson correlation coefficients were calculated for the mutational effects of that substitution with every other substitution at each position. The distribution of correlation coefficients for each substitution is shown. (D) These pairwise mutational effect score correlations are also illustrated using a force directed graph. Each node represents an amino acid and each edge force value is the Pearson correlation coefficient for the mutational effect scores of the two amino acid substitutions connected by the edge. To reduce the density of edges, only the top 40% of Pearson correlation coefficients were included. This cutoff removed proline from the graph. Amino acids are colored by physicochemical type. The graph was constructed using the networkD3 package in R.

2.2.4 *Influence of original amino acid on mutational effect*

Next, I examined whether the identity of the original amino acid affected the representativeness of different substitutions. To answer this question, I used 882 positions in with measured effects for all nineteen possible substitutions to calculate the median effect for the original amino acids, individually. Differences in the prevalence of different amino acids in the wild type protein sequences resulted in some amino acids having fewer associated mutations than others (*e.g.* leucine (N = 1,786), tryptophan (N = 114, **Fig. 2.1B**). I found that tryptophan positions were the most sensitive mutations (Trp median effect = 0.48), while glutamine positions were the least sensitive (Gln median effect = 0.99). Additionally, I observe very similar results when all positions are included in the analysis (Trp median = 0.48, Gln median = 0.99). Next, to reveal mutant amino acids that best capture the effects irrespective of original amino acid identify, I computed the difference between the median of all mutational effects observed at each original amino acid and the median mutational effects for each mutant amino acid type given the original amino acid identity (**Appendix A 2A**). Hierarchical clustering of these differences revealed two major classes of original amino acids. The first class included large hydrophobic amino acids, which are more sensitive to substitutions, while the second class included charged and polar amino acids, which are less sensitive to substitutions. Nonetheless, I found that histidine and asparagine, among other substitutions, best represent the median mutational effect for most original amino acids (**Appendix A 2B**). However, I observed that histidine and asparagine are more representative of other mutations at certain original amino acid positions. For example, histidine is least representative when the original amino acids are large, hydrophobic amino acids, tryptophan and tyrosine, or sulfur-containing amino acids, methionine and cysteine, while

asparagine mutations are least representative when the original amino acid is histidine or methionine.

2.2.5 *Correlation between mutational effect scores from different amino acids*

Because of the comprehensive nature of the large-scale mutagenesis data sets, I could ask how well the mutational effect scores of each substitution correlated with the scores of every other substitution at each position. Thus, I calculated Pearson correlation coefficients for the mutational effect scores of each substitution pair across all positions (**Fig. 2.2C, Appendix A3**). The effects of histidine and asparagine substitutions correlated best with the effects of all other substitutions, while the effect of proline substitutions correlated worst. To visualize the relationships between each pair of substitutions, I constructed a force-directed graph (**Fig. 2.2D**). As expected, substitutions cluster by physicochemical type in the graph, meaning that similar substitutions have similar effects. Proline is not represented because its effects are poorly correlated with other substitutions. Histidine and asparagine are connected to many other amino acids, owing to the high correlation of the effects of these substitutions with many other substitutions.

2.2.6 *Influence of secondary structure on patterns of mutational effects*

I next asked whether the secondary structural context of a position altered the effect of each substitution. I excluded DBR1 and GB1 from this analysis because they did not have structures of sufficiently close homologs. I used DSSP to identify 1,007 positions in the remaining proteins that were in an α-helix, a β-sheet or a turn (101). Overall, substitutions in turns are less disrupting than substitutions in α-helices or β-sheets (**Fig. 2.3A**).

However, the relative effects of each substitution in the three structural contexts were mostly consistent, especially between α-helices and β-sheets (**Fig. 2.3B, Appendix A 4A)**. Surprisingly, the tolerance for each amino acid substitution in the different secondary structural contexts was not strongly correlated with the frequency of that amino acid's occurrence in known structures (102). For example, alanine occurs more frequently in α-helices, relative to β-sheets. However, in these large-scale mutagenesis data sets, alanine substitutions were mildly disrupting in both structural contexts. These observations suggest that secondary structure does not dominate mutational tolerance, at least for the proteins I examined.

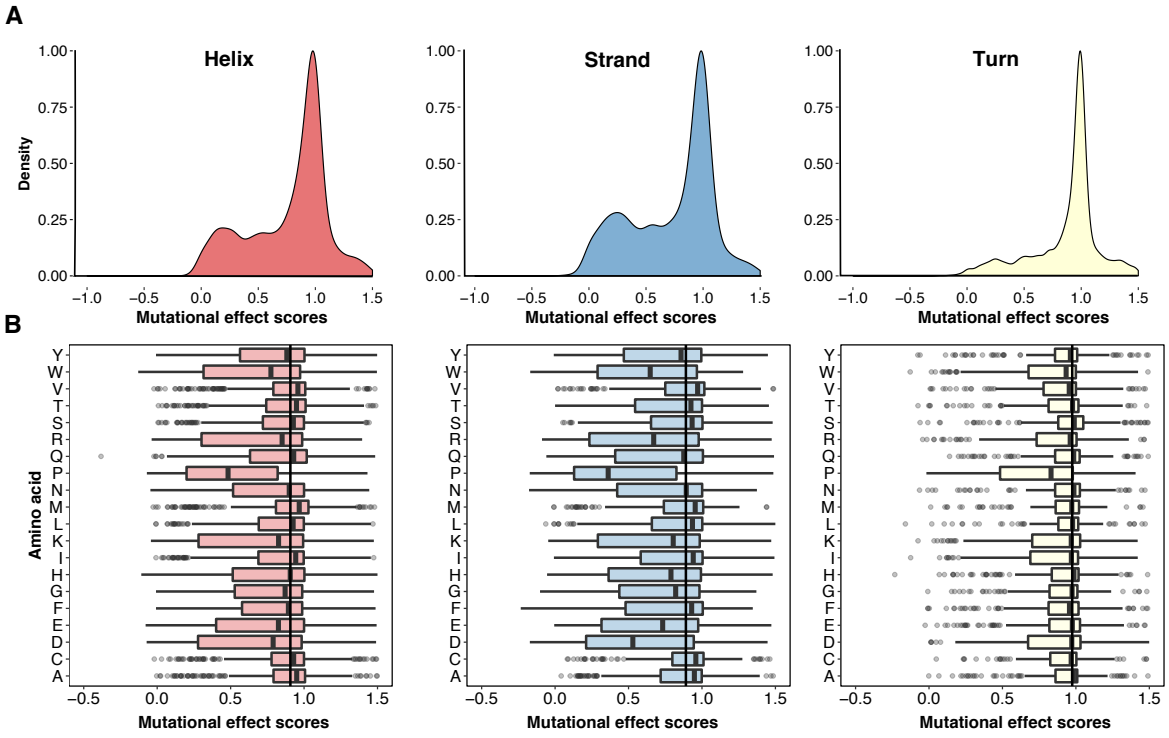


Figure 2.3. Secondary structural context of mutational effects. (**A**) Density plots describing the distribution of mutational effect scores for each substitution are shown for three different structural contexts as determined using DSSP: α-helices (left panel, N = 8,669), β-sheets (middle panel, N = 4,796), and turns (right panel, N = 3,329). (**B**) The mutational effect score distributions for each substitution inα-helices (left panel), β-sheets (middle panel), and turns (right panel) are shown. The vertical line in each panel represents the median effect score for all substitutions in that secondary structure type.

I next investigated which substitutions were the most representative regardless of structural context. I found that histidine substitutions have close to the median effect in α-helices and turns, but were more disrupting than the median effect in β-sheets (**Fig. 2.3B**). Asparagine and glutamine substitutions had near median effects in all three contexts. As above, I examined how well the effects of each substitution correlated with every other substitution at each position in each context. I found that the effects of histidine, asparagine and glutamine substitutions correlated best with the effects of other substitutions (**Appendix A 4B, C**). Thus, the effects of histidine, asparagine and glutamine are relatively consistent in the different structural contexts I examined, highlighting the representativeness of these substitutions.

2.2.7 *Predictive performance of amino acids to identify protein-ligand interfaces*

An important use of single amino acid scanning is to identify positions in protein-ligand interfaces. In order to determine whether alanine is ideal for that purpose, I analyzed the effects

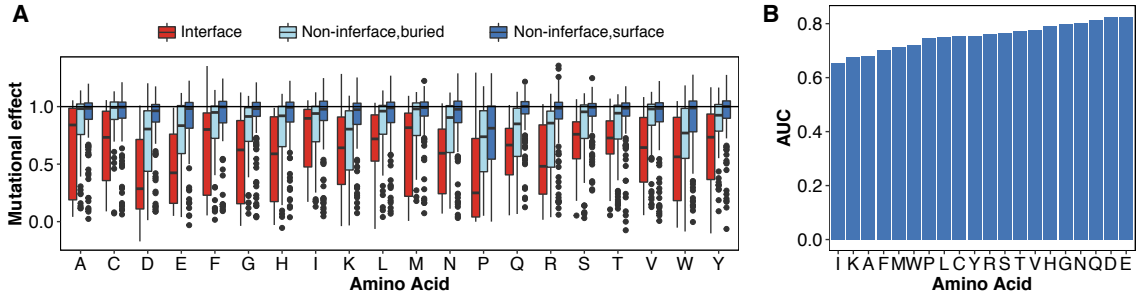


Figure 2.4 Asparagine, glutamine, aspartic acid and glutamic acid are best for identifying positions in protein-ligand interfaces. (A) The distribution of mutational effect scores for every substitution in four proteins with ligand-bound structures (hYAP65 WW domain, PSD95 pdz3 domain, BRCA1 RING domain (BARD1 binding) and Gal4) is shown at ligand interface positions as reported in the literature, and for non-interface buried positions or non-interface surface positions. (**B**) A mutational effect threshold was defined such that positions with a mutational effect below the threshold were classified as "interface," whereas positions with a mutational effect above the threshold were classified as "non-interface." ROC curves for each amino acid were generated by varying this threshold. The area under each ROC curve is shown, illustrating the power of each substitution to discriminate between interface and non-interface positions.

of substitutions in four proteins with ligand-bound structures: the hYAP65 WW domain, the PSD95 pdz3 domain, the BRCA1 RING domain and GAL4. Amongst these four proteins there were 4,884 mutations at 282 positions. I used relative solvent exposure to classify each position as either buried or on the surface. I also determined interface positions based on published structures and functional studies (see **Methods**). I found that substitutions at interface positions are substantially more disrupting than substitutions at buried, non-interface or surface, non-interface positions (**Fig. 2.4A**). This result is expected, given that all four deep mutational scans were conducted using selections that depended on ligand binding. Alanine, along with phenylalanine, isoleucine and methionine, are the least disruptive amino acid substitutions at interface positions, suggesting that they may not be ideal for interface detection.

I reasoned that the ideal substitution for detecting protein-ligand interfaces would exhibit a large difference in mutational effect between interface and non-interface positions. To formalize this idea, I used a mutational effect threshold. If a substitution at a particular position had a mutational effect below the threshold, I classified that position as "interface." Conversely, if the mutational effect was above the threshold that position was classified as "non-interface." For each substitution, I varied the mutational effect threshold from the maximum mutational effect score to the minimum effect in 200 steps. At each step, I compared the true interface positions to those determined using the mutational effect threshold procedure. I then constructed receiver operating characteristic (ROC) curves. The area under each ROC curve revealed the ability of that substitution to discriminate between true interface and non-interface positions. I found that isoleucine, lysine and alanine had among the worst discriminatory power (**Fig. 2.4B**, **Appendix A 5**). Substitutions that were highly disruptive at interfaces, like asparagine, glutamine, aspartic acid and glutamic acid, had the best discriminatory power. Next, I calculated the fraction of true

interface positions detected by each amino acid substitution at a 5% false positive rate. Here, I found that asparagine and glutamine substitutions revealed over 60% of the true interface positions; aspartic acid and glutamic acid substitutions also performed well (**Appendix A 6**). However, alanine substitutions detected fewer than 20% of the true interface positions at a 5% false positive rate. Thus, asparagine, glutamine, aspartic acid or glutamic acid substitutions are all good choices for detecting protein-ligand interfaces.

2.3 METHODS

2.3.1 *Data curation and scaling*

I curated a subset of the published deep mutational scanning data sets. I excluded deep mutational scans of non-natural proteins, because the mutational properties of natural and non-natural proteins could differ. The result was a set of sixteen deep mutational scans of fourteen proteins (**Appendix A Table 1**). BRCA1 and UBI4 each have two large-scale mutagenesis data sets corresponding independent experiments in which different functions were assayed (*e.g.* ligand binding or catalytic activity). I treated these data sets separately, and did not perform any averaging of mutational effects between the data sets. Additionally, I removed any variants with more than one amino acid substitution from all the data sets.

Most of the data sets reported mutational effect scores as the log-transformed ratio of mutant frequency before and after selection, divided by wild type frequency before and after selection. For data sets that used a different scoring scheme, I recalculated mutational effect scores as the log-transformed ratio of mutant frequency before and after selection, divided by wild type frequency before and after selection. Given that the assays used to detect mutational effect differ,

I rescaled the reported mutational effect scores for each data set. First, I subtracted the median effect of synonymous mutations from each reported effect score and then divided by the negative of the bottom 1% of reported effect scores. Finally, I added 1. In cases where synonymous mutational effect scores were unavailable, I omitted the synonymous score median subtraction step. Our rescaling scheme is expressed as

$$S_{i,scaled} = \frac{S_{i,reported} - S_{median\ synonymous}}{-S_{median\ bottom\ 1\%}} + 1$$

where S is the mutational effect score. The normalization scheme resulted in scaled mutational effect scores where the most disrupting mutations have effect scores ≈ 0 and wild-type-like mutations have scores ≈ 1. Unless otherwise stated, I used all of the rescaled mutational effect data for each analysis. In each analysis, I used median as a summary statistic rather than mean because the frequency distributions of mutational effect are bimodal rather than Gaussian (**Appendix A 1**).

2.3.2 *Variant annotation*

DSSP was used to annotate the secondary structure and absolute solvent accessibility of each wild type amino acid in the data set (http://swift.cmbi.ru.nl/gv/dssp/DSSP_3.html). To estimate the relative solvent accessibility of amino acids, I divided absolute solvent accessibility as determined using DSSP by the total surface area of each amino acid. Amino acids with relative solvent accessibilities greater than 0.2 were labeled as “surface”, whereas amino acids with relative solvent accessibilities less than 0.2 were labeled as “buried” (103).

2.3.3 *Identification of interface positions*

Four proteins in the data set had high-resolution PDB structures with peptide or nucleotide ligands, Gal4 (3COQ), BRCA1 RING domain (1JM7), PSD95 pdz3 domain (1BE9) and hYAP65 WW domain (1JMQ). I determined interface positions from the literature(5, 15, 104, 105). The interface positions in hYAP65 WW domain were 188, 190, 197 and 199. The interface positions in BRCA1 RING domain were 11, 14, 18, 93 and 96. PSD95 pdz3 domain positions were 318, 322-327, 329, 339, 372 and 379. Gal4 interface positions were 9, 15, 17, 18, 20, 21, 43, 46 and 51.

2.3.4 *Construction of ROC curves*

I constructed empirical ROC curves to illustrate the power of each substitution to discriminate between interface and non-interface positions, determined as described above. First, I defined a discrimination threshold, such that positions with a mutational effect score below the threshold were classified “interface” and positions with a mutational effect score above the threshold were classified as “non-interface.” For each substitution, I varied this discrimination threshold from the maximum mutational effect score to the minimum mutational effect score in 200 steps, calculating the true positive interface detection rate (TPR) and false positive interface detection rate (FPR) at each step. The TPR was calculated by dividing the number of interface positions with scores below the mutational effect threshold by the total number of interface positions. The FPR was calculated by dividing the number of non-interface positions with scores below the mutational effect threshold by the total number of non-interface positions. ROC curves were constructed by plotting the TPR and FPR for each of the 200 mutational effect thresholds. The

area under each ROC curve was determined in R using the auc() function in the pROC package (https://cran.r-project.org/web/packages/pROC/pROC.pdf).

2.4 Discussion

Single amino acid scanning mutagenesis is a widely-used method for identifying protein positions that are important for function or ligand binding. Alanine is often employed, and was selected on rational grounds, as it constitutes a deletion of the side chain at the β-carbon. By analyzing tens of thousands of mutations in fourteen proteins, I have determined that alanine is not the most revealing substitution. For example, histidine and asparagine substitutions have an effect close to the median, and these substitutions correlate best with the effects of all other substitutions. Thus, they better represent the effects of mutations generally. Asparagine, glutamine, aspartic acid and glutamic acid are the most useful substitutions for detecting ligand interface positions. Thus, this work highlights the utility of large-scale mutagenesis data and suggests that alanine is not necessarily the best choice for future single substitution mutational scans whose goal is to identify functionally important positions or map protein-ligand interfaces.

However, these conclusions are based on only fourteen proteins. While these proteins are diverse in structure and function, they may not fully reflect the mutational propensities of other proteins. For example, tryptophan scanning mutagenesis is often applied to transmembrane domains (106-108), which were absent from the proteins I analyzed. Thus, these conclusions are most applicable to soluble proteins. Furthermore, I do not address specialized applications of single amino acid scanning mutagenesis. For example, cysteine scanning mutagenesis has been used to introduce disulfide bridges (88) and glycine scanning mutagenesis has been used to increase

conformational flexibility (109). These conclusions do not apply to these situations. Finally, the deep mutational scanning data I analyzed arises from genetic selections for protein function. Biochemical assays might reveal different patterns. However, I note that a few of the large-scale mutagenesis data sets I used were benchmarked against and found to be consistent with biochemical assay results (110, 111).

Deep mutational scanning can reveal the functional consequences of all possible single amino acid substitutions in a protein. However, these experiments can be expensive or unwieldy. Therefore, scanning mutagenesis with one or a few amino acids will remain useful for determining functionally important positions, probing protein-ligand interactions and answering other specific questions. These results could be used to guide future single amino acid scanning mutagenesis experiments, enabling selection of the amino acid best suited for the goals of the experiment.

Chapter 3. QUANTITATIVE MISSENSE VARIANT EFFECT PREDICTION USING LARGE-SCALE MUTAGENESIS DATA

3.1 INTRODUCTION

Mutations have the power to completely reshape protein structure, stability or activity and can have drastic effects on evolutionary fitness, protein function and human health. For example, mutations were used to improve the pharmacokinetic and pharmacodynamic properties of insulin to more effectively treat diabetes (112). Moreover, a recent survey of genetic variation in humans revealed that each individual harbors ~50 private missense variants, most of which are of unknown effect (113, 114). This example highlights an important trend: DNA sequencing advances have facilitated detection of genetic variation. However, in both laboratory and clinical settings, determining the impact of a missense variant on a protein's function remains a challenge (115).

Experiments can reveal a variant's molecular effect, and recent advances in multiplex assays have enabled the assessment of large numbers of variants (4, 116). However, I remain far from having a comprehensive atlas of missense variant effects in the human proteome, and such an atlas for commonly studied model organisms is also a distant goal. Thus, variant effect predictors such as PolyPhen2 (117), SIFT (34), SNAP2 (42), Evolutionary Action (38), CADD (40) and a host of others (48) will continue to be widely used to predict missense variant effects. Some predictors are products of sophisticated supervised machine learning algorithms, and are developed using features and training data that make them suited for a particular type of prediction problem. For instance, the PolyPhen2 HumDiv model is a support vector machine

trained on thousands of human Mendelian disease-associated and neutral variants from the Swiss-Prot database, and is thus optimized to predict the clinical effect of human variants (117). SNAP2, an ensemble of neural network models, is trained on human pathogenic and neutral variants as well as variants that impact molecular function (42). Given the breadth of training data, SNAP2 predictions encompass both the clinical and molecular effects of missense variants. Conversely, SIFT and Evolutionary Action are not products of machine learning, but instead rely on evolutionary patterns to predict variant effects. Despite their simplicity, SIFT and Evolutionary Action perform similarly to PolyPhen2 and SNAP2 for various prediction tasks (38), which highlights the importance of evolutionary information to successful variant effect prediction. A recently described unsupervised method, EVmutation, leverages evolutionary signatures of epistasis to predict variant effects, and has demonstrated enhanced predictive accuracy over SIFT and PolyPhen2 for both molecular and clinical effect prediction (49). These tools are all used to prioritize variants in clinical and laboratory settings.

Current predictors face two major limitations. First, most are optimized to predict categorical variant effects (*e.g.* damaging vs. benign), and cannot accurately predict effect magnitude. This limitation arises primarily from the structure of variant effect databases used to train predictors. For example, the Human Gene Mutation Database (118), Online Mendelian Inheritance of Man (119), and ClinVar (120) all categorize variants as deleterious or benign to human health. Swiss-Prot and the Protein Mutational Database contain categorical measures of variant effects in laboratory assays. Second, most predictors focus on predicting the clinical effect of human variants rather than the molecular effects on protein function (34, 117). However, the relationship between molecular effect and clinical effect is complex, and most predictors do not

deal well with this complexity. For example, both gain- and loss-of-function variants of BRAF can be pathogenic (121, 122). Variants of PTEN can drive carcinogenesis when they occur somatically, or can cause autism or a tumor syndrome when they occur in the germline (123). Thus, I suggest that accurate clinical effect prediction should start with accurate, quantitative predictions of molecular effect whose subsequent interpretation is guided by specific knowledge about gene-disease associations.

Here I address the need for an accurate, quantitative predictor of molecular effect by leveraging deep mutational scanning data. In a deep mutational scan, selection for protein function among a library of nearly all possible single amino acid variants of a protein is coupled to high-throughput DNA sequencing (4, 21). Sequencing reveals how each variant's frequency changes during selection, and these changes provide quantitative scores that describe the functional effect of each variant in the library. The resulting large-scale mutagenesis datasets have a distinct advantage over traditionally-used variant effect predictor training datasets like HumDiv/HumVar, HGMD and the Protein Mutant Database. Traditional datasets contain a large number of proteins, each with a median of four to six variant effect measurements. A large-scale mutagenesis dataset contains deep and unbiased information, capturing the effects of most variants at every position in single protein. I hypothesize that large-scale mutagenesis datasets contain informative and generalizable patterns that can be used to predict variant effects in disparate proteins.

Mutations have the power to completely reshape protein structure, stability or activity and can have drastic effects on evolutionary fitness, protein function and human health. For example,

mutations were used to improve the pharmacokinetic and pharmacodynamic properties of insulin to more effectively treat diabetes (112). Moreover, a recent survey of genetic variation in humans revealed that each individual harbors ~50 private missense variants, most of which are of unknown effect. This example highlights an important trend: DNA sequencing advances have facilitated detection of genetic variation. However, in both laboratory and clinical settings, determining the impact of a missense variant on a protein's function remains a challenge.

Here I address the need for an accurate, quantitative predictor of molecular effect by leveraging deep mutational scanning data. In a deep mutational scan, selection for protein function among a library of nearly all possible single amino acid variants of a protein is coupled to high-throughput DNA sequencing (5, 20). Sequencing reveals how each variant's frequency changes during selection, and these changes provide quantitative scores that describe the functional effect of each variant in the library. I use the molecular effects of 21,026 variants of eight proteins, determined through deep mutational scans, to train Envision, a decision tree ensemble-based quantitative variant effect predictor. Envision uses a stochastic gradient boosting learning algorithm, which excels at analyzing nonlinear interactions between features and has performed well in a myriad of regression tasks (51). To maximize Envision's generalizability, proteins in the Envision training set have disparate structures and functions, and are drawn from diverse organisms. I demonstrate the generality of Envision's predictions by iteratively training models that exclude a single protein dataset and then comparing the resulting model's predictions to the observed variant effects for the excluded protein. I also assess performance using independent variant effect data that was not generated by deep mutational scanning nor included in Envision's training. I observe that Envision's predictions are generally more accurate than other state-of-the

art predictors. Envision's prediction accuracy is also highly consistent across different amino acids, unlike other predictors that perform well on some amino acids and poorly on others. I pre-computed Envision predictions for all possible single amino acid variants of proteins in the human, mouse, fruit fly, clawed frog, zebrafish, worm, and yeast proteomes. I provide a web-based tool allowing users to visualize and explore predicted protein sequence-function maps, which can be used to guide experimental assays aimed at studying variant effect. Envision is available at https://envision.gs.washington.edu.

3.2 Results

3.2.1 *Data Collection and Curation*

I collected previously published, large-scale mutagenesis datasets with quantitative measures of variant effect on protein function. Exploratory analysis led to the following criteria for inclusion: 1) the experiment must have measured single amino acid variant effects, rather than averaging effects across different genetic backgrounds; 2) the experiment

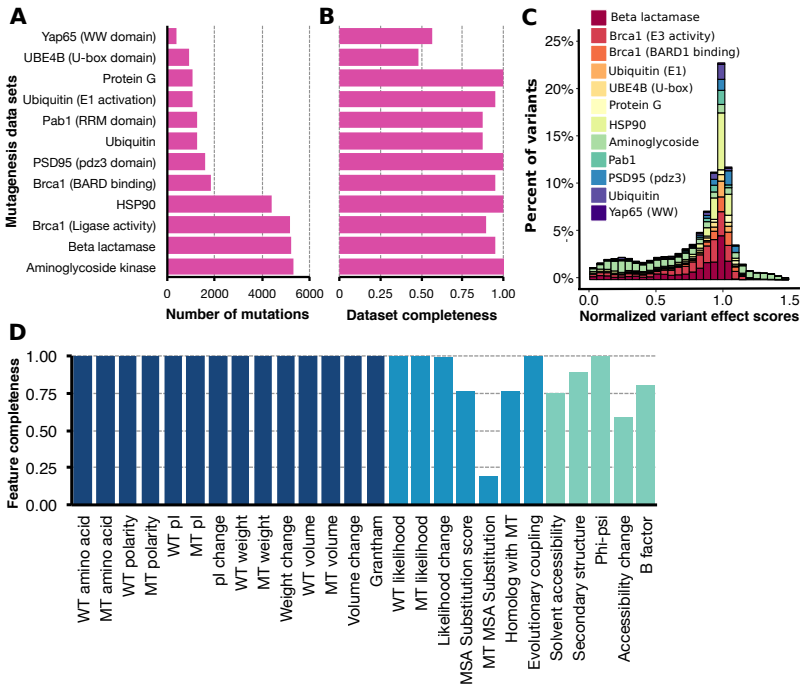


Figure 3-1. Large-scale mutagenesis data and descriptive features used to train Envision. The number of single mutants (**A**) collected from different protein or protein domain large-scale mutagenesis datasets and the mutational completeness of each dataset (**B**) are shown. Mutational completeness was calculated by dividing the number of observed single mutants by the number possible single mutants. (**C**) The distribution of variant effect scores for each large-scale mutagenesis dataset is shown. For each dataset, variant effect scores were normalized such that a score of one is wild type-like and a score of zero is inactivating. Each collected variant was annotated with 27 features, which describe physicochemical (dark blue), evolutionary (blue) or structural (green) variant attributes. (**D**) The proportion of variants in the collected large-scale mutagenesis datasets having each feature is shown (WT = wild type, MT = mutant).

must have been on a natural protein instead of a designed protein; and 3) the experiment must have quantitated molecular effects for at least ~50% of all possible variants within the mutagenized region. Ultimately, deep mutational scans of ten proteins from twelve studies comprising 28,545 single amino acid variant effects met these criteria (**Figure 2.1A, Supplementary Table 1).** Variant coverage ranged from ~50% for the Ube4b domain of murine E3 ligase to 100% for the IgG-binding domain of influenza protein G and the PDZ domain of human PSD-95 (**Figure 3.1B**). Variant coverage depended on experimental details like the protocol used for library generation (e.g., doped oligomer (124) vs. site saturation mutagenesis (8)), the number of clones generated and the depth of sequencing. The proteins in the dataset were distinct, coming from different organisms, having different structural folds and serving molecular functions ranging from catalysis to peptide binding (**Appendix B Table 1**). To make datasets comparable to one another, I normalized variant effect scores in each dataset such that variants that were more active than wild-type had a variant effect score greater than one, wild-type-like variants had a score of one and variants that were less active than wild-type had a score less than one (**Appendix B 1, Fig. 3.1C**).

Next, I annotated each variant with 27 descriptive biological, structural and physicochemical features. The biological features captured evolutionary constraints using both site-specific and co-varying conservation metrics. The structural features included local density and solvent accessibility from DSSP24, while the physicochemical features describe properties of amino acids, such as polarity and size. Physicochemical and biological features were available for nearly all variants, but structural features were not (**Figure 3.1D, Appendix B Table 2**).

3.2.2 *Predicting Quantitative Variant Effects*

I first tested whether a stochastic gradient boosting regression algorithm could model the relationships between these 27 descriptive features and quantitative variant effect scores for each protein, individually. To train each single-protein model, hyperparameters, such as the number of decision trees in the ensemble and tree depth were tuned using tenfold cross-validation (see Methods). After hyperparameter tuning, I reserved a distinct subset of mutations for testing by limiting training to a random 80% of each large-scale mutagenesis dataset. This approach allowed us to estimate the generality of each model to unseen variants within each protein. Nine of the twelve models performed very well (median Pearson's R = 0.83, Spearman's ρ = 0.80, Figure 2.2A), while three, the BRCA1 RING domain BARD binding, BRCA1 RING domain E3 activity and E4B ubiquitin ligase models, performed poorly (median R = 0.22, ρ = 0.35).

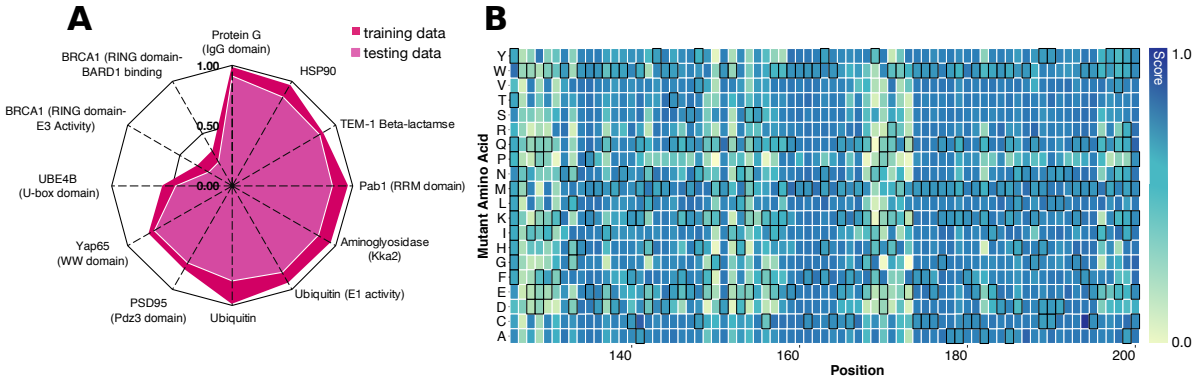


Figure 3-2 Protein-specific gradient boosting models can accurately predict variant effect scores. I trained a model for each protein using a randomly selected 80% of data, with 20% reserved for testing. (**A**) A radar plot of Pearson's correlation coefficients between observed and predicted variant effect scores illustrates protein-specific model performance on both training (dark red) and testing data (light red). The PAB1 RRM domain-specific model predicts the effects of variants withheld from training well (Pearson's R > 0.75), and was used to predict the 197 missing variant effect scores. (**B**) The completed Pab1 RRM domain sequence-function map is shown for positions 126-200. Each mutagenized position is a column, and each amino acid substitution is a row. Wild type-like variants are colored dark blue and inactive variants are colored light blue. Predicted effects are denoted by black borders.

Experimental noise cannot account for these models' poor performance, since the correlation of model predictions with the training and testing data is much lower than the correlation between replicate experiments (**Appendix B Table 1**). I hypothesized that poor performance arose because correlations between the features and variant effect scores were low (**Appendix B 2**). The low correlation might occur because the assays did not test natural functions of these proteins, and, whereas some of the model features are biochemical, many are evolutionary. For instance, the BCRA1 RING domain variants were assayed for two specific functions, E3 ligase activity and BARD binding. However, BRCA1 is a complex protein with many functions and interactions with more than 25 other proteins (16). Another possibility is that these two datasets were missing some structural features. However, the YAP65 WW domain dataset was missing the same features yet resulted in an accurate model. Thus, I cannot definitively identify the cause of poor performance in the BRCA1 RING domain and E4B ubiquitin ligase models. I excluded these three datasets from subsequent analyses.

For most proteins, the feature set and learning procedure generated accurate models of variant effect. Beyond validating our approach, these single protein models afforded us the opportunity to complete each large-scale mutagenesis dataset by predicting missing variant effect scores. The nine datasets I analyzed were collectively missing 862 variants, or 4% of the possible single amino acid variants. The Pab1 dataset lacked effect scores for ~20% of the possible variants. I used the Pab1 model ($R = 0.86$; $\rho = 0.79$) to predict the missing data and complete the Pab1 RNA recognition motif dataset (**Figure 3.2B**). I provide completed datasets for each protein (**Appendix B Table 3**).

Next, I used stochastic gradient boosting to train a global model with the 21,026 empirically-derived variant effect scores in the nine large-scale mutagenesis datasets (see Methods). I crafted a cross-validation scheme specifically designed to avoid protein-specific overtraining. In this scheme, I tuned hyperparameters using a leave-one-protein-out approach. During each round of cross-validation, I withheld all variant effect scores from one of the proteins, training only on variant effect scores from the other seven proteins, and then predicted the held-out dataset in the validation phase (**Appendix B 3A, Appendix B Table 4**). Once hyperparameters were tuned, I trained Envision with all available data, except for a random 5% of variant effect scores that I withheld for testing and to assess overfitting. Training and testing data root mean squared errors were similar at each model training iteration, indicating that the model is not overfitted to the training data (**Appendix B 3B**). Envision predicted the training data well ($R = 0.79$, $\rho = 0.76$; Figure 2.3A).

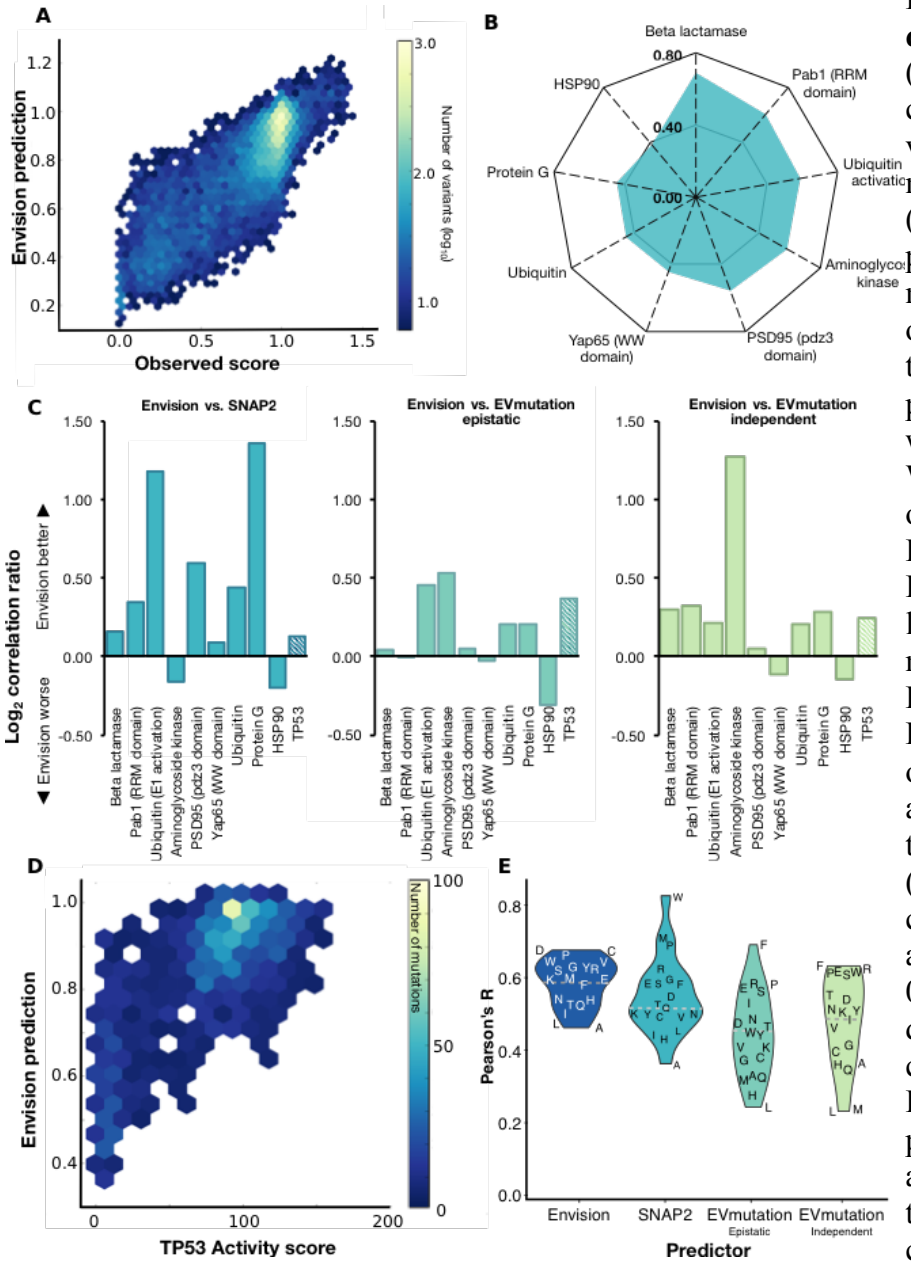


Figure 3-3 Envision outperforms other quantitative variant effect predictors. (**A**) A hexagonal bin plot shows the correlation between predicted and observed variant effect scores for all the large-scale mutagenesis data used to train Envision (Pearson's R = 0.79). To evaluate performance on data not used in training, models were retrained excluding each one of the nine proteins. (**B**) A radar plot shows the correlation (Pearson's R) between predicted and observed variant effect scores when the indicated protein was left out. (**C**) We also compared the leave-one-protein-out models to SNAP2 (left panel), EVmutation-epistatic (middle panel) and EVmutation-independent (right panel). The $\log_2$ ratio of each leave-one-protein-out model's Pearson's R to another predictor Pearson's R on the left-out data is shown. Hashed bars indicate relative performance on a set of 2,312 TP53 transactivation activity scores measured in a low-throughput assay and not used in training. (**D**) A hexagonal bin plot shows the correlation between Envision predictions and TP53 activity scores (Pearson's R = 0.58). (**E**) A violin plot illustrates the distribution of Pearson's correlation coefficients for variant effect scores and Envision, SNAP2 and EVmutation predictions for different mutant amino acids. The dashed horizontal line indicates the median Pearson's correlation coefficients for each predictor.

3.2.3 *Assessing Envision Performance*

To evaluate Envision's performance, I employed a jack-knife approach to estimate how well a leave-one-protein-out (LOPO) model could predict variant effect scores for a protein excluded

from model training. Here, I repeated the training procedure described above, leaving one protein completely out of the hyperparameter tuning and model training process. Then, I used the resulting model to predict variant effect scores for the left-out protein and determined performance (see Methods). I repeated this procedure for all nine proteins. Variant effect scores for left-out proteins were predicted with Pearson's R ranging from 0.38 to 0.69 and Spearman's ρ ranging from 0.30 to 0.74 (**Figure 3.3B; Appendix B 3C**). To determine the effect of the variant effect score normalization scheme on model training and performance, I compared LOPO models trained using either normalized or non-normalized variant effect scores. Indeed, models trained using normalized data predicted variant effect scores for the left-out protein better than models trained using non-normalized data (median R = 0.56 vs. 0.39, median ρ = 0.51 vs. 0.35; **Appendix B 3D**). This result highlights the utility of the variant effect score normalization scheme.

Next, I compared our LOPO models' performance to other predictors. PolyPhen2 is trained to predict the categorical clinical effect of variants, but also generates a numerical score. This score is the naïve Bayes posterior probability that a variant is damaging, and, although quantitative, it is not designed to predict the magnitude of a variant's molecular effect. As expected, for the only human protein in our dataset, YAP65 WW domain, our LOPO model outperformed PolyPhen2 when predicting WW domain variant effect scores (R = 0.46 vs. 0.17; ρ = 0.36 vs. 0.19). Like PolyPhen2, SIFT also generates categorical predictions and scores for human proteins. SIFT scores represent the scaled probability of a missense variant being tolerated, and are also not expected to capture the magnitude of variant molecular effects. The WW domain LOPO model also outperformed SIFT scores (R 0.46 vs. 0.03; ρ = 0.36 vs. 0.04). PolyPhen2 and SIFT were

not designed to predict variant effect magnitude, and these results confirm that they should not be used to do so.

SNAP2, EVmutation and Evolutionary Action were developed to predict variant effect magnitude(38, 42, 49). However, Evolutionary Action scores could not be obtained by batch query, preventing us from including them in this analysis. I found that SNAP2 predicted variant effect scores much better than PolyPhen2 or SIFT, but not as well as our LOPO models, which outperformed SNAP2 on seven of nine datasets (median R 0.56 vs. 0.44) (**Figure 3.3C**). I also compared our models to EVmutation, which predicted the magnitude of variant effects using either an epistatic or an independent conservation-based unsupervised statistical model. Here, our LOPO models outperformed EVmutation's epistatic model on six out of nine datasets (median R 0.56 vs. 0.47) and EVmutation's independent model on seven of nine (mean R 0.56 vs. 0.48; **Figure 3.3C**). An equivalent analysis using Spearman's ρ revealed similar results (**Appendix B 3E**). I also assessed the magnitude of improvement of our LOPO models over other tools in terms of Pearson R. Across all datasets, our LOPO models' predictions are 4%, 14% and 21% more correlated with the observed variant effect scores than predictions from EVmutation epistatic, EVmutation independent and SNAP2 models, respectively.

Next, I aimed to determine what factors led to our improved performance. Many differences exist between Envision and the other tools, making it difficult to unequivocally identify the sources of improvement. For example, Envision, SIFT, EVmutation, PolyPhen2 and SNAP2 were all generated using different algorithms that can't readily be compared. Envision's features are similar to those used by SNAP2 and PolyPhen2, so the improvement I observed is not likely due to Envision's features. Instead, I hypothesized that our use of deep mutational scanning data

and our cross-validation approach, designed to yield a generalizable model, are the two attributes that led to improved performance. The lack of a large database of quantitative variant effects measured by means other than deep mutational scanning made it impossible to evaluate the performance advantage conferred by training our models using deep mutational scanning data. However, I could quantify the impact of our cross-validation approach by comparing the performance of models trained using a standard tenfold cross-validation hyperparameter tuning scheme to models trained using our leave-one-protein-out (LOPO) scheme. I found that hyperparameter tuning using our LOPO approach yielded improved performance compared to tenfold cross-validation (median $R = 0.56$ vs. 0.45, $\rho = 0.50$ vs. 0.45; **Appendix B 3F**). Thus, our LOPO cross validation procedure, which improved predictive performance by ~10-20% over all protein datasets, is a cornerstone of Envision's success. I suggest that our LOPO approach, designed to yield generalizable models, was especially important given that our training data set contained relatively few proteins.

Our leave-one-protein-out analysis demonstrated that Envision provided improved quantitative predictions of variant effect for variant effect scores measured using deep mutational scanning. However, I was concerned that Envision's performance advantage arose because it learned deep mutational scanning-specific patterns in the data. To ensure that Envision was not overfitted to data derived from these methods, I obtained a mutagenesis dataset for the TP53 tumor suppressor that was not generating using deep mutational scanning. Instead, the effect of each variant on TP53 transactivation was measured individually in a plate-based assay. Here, TP53 variant transactivation of eight distinct TP53 response-elements was measured using a fluorescent reporter protein (125). Overall, this dataset contained 2,312 TP53 variant effects expressed as a

percentage of wild-type fluorescence, averaged across the eight response elements. I predicted TP53 variant effect scores using Envision, which was trained on all nine large-scale mutagenesis data sets. The TP53 data were never used, directly or indirectly, during the training procedure. Despite the fact that the TP53 dataset was not acquired using deep mutational scanning, Envision predicted the TP53 variant effect scores well ($R = 0.58$, $\rho = 0.53$; **Figure 3.3C, D**). Importantly, Envision outperformed SNAP2 ($R = 0.53$; $\rho = 0.50$), whose training dataset included ~400 human TP53 mutations from more than 50 independent studies, and EVmutation (epistatic $R = 0.45$, $\rho = 0.49$; independent $R = 0.49$, $\rho = 0.52$; **Figure 3.3C**). Thus, Envision has learned patterns of the molecular effects of variants that do not appear to depend on the measurement method.

Next, I sought to determine whether Envision performance depended on the identity of either the mutant or wild type amino acid. I evaluated performance on the TP53 dataset to enable comparison to EVmutation and SNAP2. I found that Envision prediction performance did not depend much on the identity of the mutant amino acid (**Figure 3.3E, Appendix B 3G**). However, EVmutation and SNAP2 showed large biases in performance. For instance, EVmutation predicted mutations to phenylalanine with high accuracy ($R = 0.69$, $\rho = 0.70$), but predicted mutations to leucine with low accuracy ($R = 0.24$, $\rho = 0.33$). SNAP2 performance was also biased in favor of mutations to tryptophan and methionine and against mutations to alanine. These biases are also apparent in the context of the wild-type amino acid, where I found EVmutation predicted mutations from wild type cysteine very well ($R = 0.82$, $\rho = 0.71$) and wild type aspartic acid poorly ($R = 0.02$, $\rho = 0.05$; **Appendix B 3H**). Consequently, in addition to greater overall accuracy, Envision performance was more consistent.

Finally, I assessed the utility of Envision scores for clinical effect prediction. I evaluated performance by constructing ROC curves using variants annotated as either pathogenic or benign in the ClinVar database. Envision predictions were much better than random chance (AUROC = 0.72), but not as good as PolyPhen2, CADD and SIFT (AUROCs = 0.86, 0.85, 0.84; **Appendix B 3I**). This result is not surprising because Envision was not designed or optimized for this task, and because comparison of predictor performance on clinical data is difficult given that many predictors are trained on or optimized to predict these data28. Furthermore, the relationship between the magnitude of a variant's molecular effect and disease phenotype is likely to be different for each disease-associated protein. For example, a weakly damaging variant in some proteins may be sufficient to cause disease, whereas only strongly damaging variants lead to disease in other proteins. Finally, I note that the rate at which training datasets grow in the coming years may be much greater for deep mutational scans than for clinical variant databases.

3.2.4 *Feature importance and future improvements*

Gradient boosting models are highly interpretable. By analyzing which features were commonly used in the ensemble of decision trees, I estimated the importance of each feature to deepen our understanding of the properties of amino acid variants that alter function. The features that were most strongly represented in the Envision decision tree ensemble include structural features like B factor and solvent accessibility, as well as biological features related to evolutionary conservation (**Figure 3.4A; Appendix B Table 5**). These features are known to be predictive of variant effects (126, 127). However, unlike other feature-driven predictors (42, 117), I found that the mutant amino acid identity is informative. This amino acid identity effect was largely driven

by proline. Proline variants are generally disruptive of protein function because they eliminate the amide hydrogen necessary for hydrogen bonding and because they constrain the conformation of the polypeptide backbone. Indeed, proline variants were more damaging than other substitutions in the large-scale mutagenesis datasets (proline mean effect score = 0.60 vs. all AA mean = 0.81; paired t-test $P << 0.001$, n = 8; **Appendix B Figure 9**). Additionally, I found that Envision predicted the effects of proline variants about as accurately as the effects of other variants (**Appendix B Figure 10**). Thus, rather than simply predicting that all proline variants were strongly damaging, Envision predicted the degree to which proline variants maintain or disrupt function.

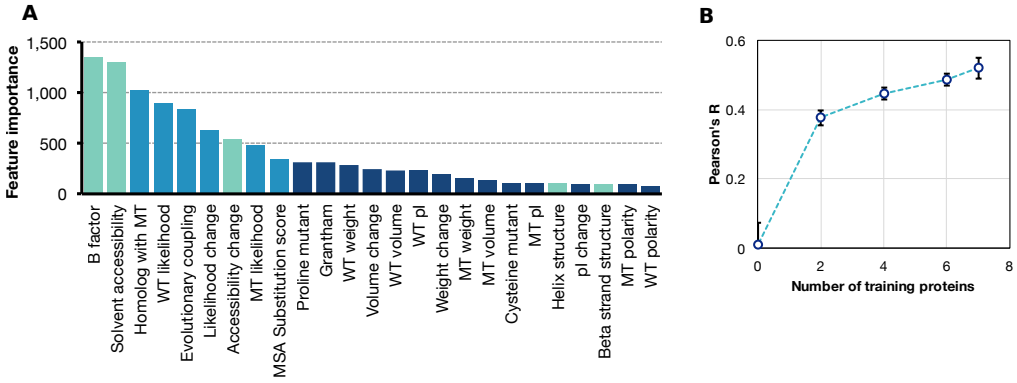


Figure 3-4 Envision is an interpretable model that will improve with more training data. The number of times each feature is used in Envision's decision tree ensemble is a measure of feature importance. (**A**) Feature importance for every physicochemical (dark blue), biological (blue) and structural (green) is shown (WT = wild type, MT = mutant). To assess the impact of adding more training data to Envision, I conducted a downsampling analysis. Models were trained with increasing numbers of randomly selected protein datasets, and tested on mutations from proteins withheld from training. (**B**) The mean Pearson's correlation coefficient between predicted and observed variant effects across testing datasets are shown, organized by the number of proteins included in the training set. Error bars indicate the standard deviation of correlation coeffcients obtained from ten random samplings of proteins to include in the training set. A naïve model (*i.e.* number of training proteins = 0) was also generated by randomizing feature values for all proteins and repeating the training procedure. The error bars for the naïve model indicate the standard deviation of correlation coefficients obtained from ten different feature randomizations.

Finally, I hypothesized that incorporating additional large-scale mutagenesis datasets into future versions of Envision would improve performance. To determine how the number of proteins in our training dataset affected the generality of Envision to unseen data, I performed a down-sampling analysis. Here, versions of Envision were trained with datasets containing different numbers of proteins and tested on withheld protein datasets. I found that model generality increased as more proteins were used in model training, suggesting that accumulation of more data will improve Envision's predictive performance (**Figure 3.4B**).

3.3 Discussion

I developed Envision, the first variant molecular effect predictor trained on large-scale mutagenesis data. Envision is generalizable, accurately predicting variant effects in large-scale mutagenesis data withheld from training. Envision can also accurately predict variant effects from low-throughput experiments, and thus has not simply learned underlying patterns exclusive to large-scale mutagenesis data. Overall, Envision outperforms other quantitative predictors like SNAP2 and EVmutation in predicting experimentally measured molecular effects. In particular, the quality of Envision predictions is relatively uniform across different amino acid substitutions, whereas other predictors' accuracy is driven by high performance on some substitutions and poor performance on others. The promise of using large-scale mutagenesis data to develop variant effect predictors is highlighted by the fact that Envision was trained from deep mutational data on only nine proteins, but can outperform established methods that are trained using sparse mutational data on thousands of proteins. Thus, as more large-scale mutagenesis data becomes available, Envision will continue to improve.

Envision also has limitations, which may hinder its utility. Envision's predictions are provided as quantitative scores that range from ~0 to ~1, where scores less than one are function-damaging as compared to wild-type. I find that Envision can predict the scores of strongly damaging and neutral mutations well, but predicts mutations of intermediate effect less well (**Appendix B 3C**). Envision also relies on structural and evolutionary features that are not available for every protein, and I found that predictive performance degraded when these features were missing. Thus, while Envision predictions are available for millions of variants, I recommend treating them with caution when key features are missing. Consequently, the Envision web tool reveals which features are incorporated for each prediction.

Providing evidence for or against human variant pathogenicity is an important problem, made challenging by the complex relationship between molecular and organismal effect. In some cases, like cystic fibrosis, significant loss of function is required to cause disease (128). In other cases, even minor alterations are pathogenic (129). Proteins have many functions, many interaction partners and can be related to many diseases. An ideal pathogenicity predictor would incorporate accurate, quantitative predictions of molecular effect with gene-disease association and population variation data. Because Envision is trained on, and makes predictions of, molecular effects, it could be useful in developing the next generation of pathogenicity predictors.

I furnish pre-computed Envision predictions for all possible single amino acid variants of each gene for six model organism and human proteomes. I anticipate that Envision will be a useful tool for identifying candidate variants that tune protein activity levels. Such variants may serve

as a mechanism to control protein activity in *in vitro* or *in vivo* systems. Envision's predictions of molecular effect may also be useful in situations where the relationship between protein function and disease is clear. Thus, Envision can be useful in a number of ways, and will continue to improve as new datasets become available.

3.4 METHODS

3.4.1 *Training and data collection*

Published large-scale mutagenesis datasets were used as training data if they met several criteria. First, I required that at least 50% of all possible single amino acids were substituted at each mutagenized position. Thus, alanine and proline scans did not qualify for this study. Second, I only accepted mutational scans of native proteins assayed for native biological function. Third, I excluded scans in which the complete variant sequence was unknown. I also removed variants with more than one mutation. In total, I accepted twelve datasets comprising ~30,000 missense mutations. These scans were performed on proteins from different organisms: human, mouse, rat, *S. cerevisiae*, and bacteria (**Appendix B Table 1**).

3.4.2 *Normalization*

Each large-scale mutagenesis dataset was generated using a distinct experimental assay, which resulted in different variant effect score distributions. To enable meaningful comparison between datasets, I normalized them. For each dataset, every variant effect score was normalized to the wild type score and then $\log_2$ transformed (**Appendix B 1A**). Next, I subtract the median effect of synonymous variants, if available. Synonymous variants were unavailable for the PSD95

(Pdz3 domain), Protein G (IgG domain), UBE4B (U-box domain) and BRCA1 datasets, so I instead subtracted 0 from each score in those datasets. Lastly, I divided each score by the negative median score of the bottom 1% of mutations of each dataset and added one. Our normalization scheme is expressed as

$$S_{normalized} = (S_{reported\ i} - S_{median\ synonymous})/(-S_{median\ bottom\ 1\%}) + 1, \quad (2.1)$$

where S signifies score. This normalization scheme results in variants that are more active than wild type having scores of greater than one, wild type-like variants having scores of one, and damaging variants having scores of less than one (**Appendix B 1B**).

3.4.3 *Variant annotation*

Mutations were annotated with three general types of descriptive annotations: evolutionary, biochemical and structural (**Appendix B Table 2**). Several evolutionary features used in our model were obtained using the PolyPhen2 annotation pipeline (130). I also derived a measure of average mutational covariance between a given position and all other positions in a multiple sequence alignment from EVfold (49). To obtain structural information, I use DSSP (http://www.cmbi.ru.nl/dssp.html) (101) and PDB files from the Protein Data Bank (http://www.rcsb.org/pdb/home/home.do) (131). Our biochemical annotations include measures of amino acid size, weight, volume, isoelectric point, and Grantham scores (100).

3.4.4 *Machine learning*

Stochastic gradient boosting is a method of machine learning that uses an ensemble of weak prediction models (*e.g.,* decision trees) for classification or regression problems (51). I constructed stochastic gradient boosting tree regression models using the *GraphLab Create* framework from Turi (https://turi.com/products/create/). Hyperparameters were optimized using a grid search. For each predictive model, I tuned six parameters in a stepwise fashion. First, I optimized for the number of decision trees in the ensemble. Next, I tuned the maximum depth of a decision tree and the minimum number of observances allowed in a terminal node of a tree. Then, I determined the value that the squared-loss must be reduced by in order to add an additional node to a tree. Finally, I identified the optimal proportion of variant effect scores and features used to train each tree. Once hyperparameters were tuned, I reduced the learning rate from 0.1 to 0.01 and increased the number of decision trees by five-fold.

3.4.5 *Single protein models*

To filter out datasets that are noisy or contain variant effects that cannot be explained by our evolutionary, structural or physicochemical features, I performed gradient boosting machine learning on a randomly selected 80% of variant effect scores from each protein dataset. This resulted in a model for each protein, which I used to predict the 20% of variant effect scores withheld from model training. Proteins whose specific models performed poorly on withheld data (Pearson's $R < 0.5$) were excluded from the LOPO and global models.

3.4.6 *Training Envision*

Envision was trained using the same approach as our single protein models with an added leave-one-protein-out cross-validation procedure, where, at each round, a different protein was removed from the training set and used for validation (**Appendix B 3**). Thus, after each round of training, a model's generality was tested on variant effect scores from a protein not used to train the model. This cross-validation procedure allowed us to test an array of hyperparameters to see which parameter sets yielded the most generalizable models. Here, model generality was determined by measuring the root mean squared error between model predictions and variant effect scores from a left-out protein. Once all hyperparameters were optimized (**Appendix B Table 4**), I trained *Envision* with all available data except for a randomly selected 5% of which I excluded to evaluate model generality and ensure that the model was not overfitted.

3.4.7 *Leave-one-protein-out (LOPO) models*

To estimate Envision's performance on proteins not used in model training, I generated nine LOPO models. These models were trained using the same protocol as Envision, except that in each case a different protein was left completely out of the hyperparameter tuning and final model training procedures.

3.4.8 *Downsampling analysis*

To evaluate the effect of additional training data on model performance, I trained models with 2, 4, 6 or 7 of the available nine protein datasets. Model training was performed as described above. Each model was used to predict variant effects in proteins that were not used during the training

phase. Confidence intervals were generated by randomly selected proteins to use in the training phase.

Chapter 4. EFFECT OF MUTATION ON PROTEIN AGGREGATION

4.1 INTRODUCTION

Protein aggregation is a molecular phenomenon that affects all known organisms from bacteria to humans and implicated in a number of human diseases (132). One aggregation-associated disease is Alzheimer's disease (AD), which is the most common form of dementia. AD is incurable, untreatable and unpreventable, as its molecular underpinnings remain unknown. However, genetic, biochemical and epidemiological evidence suggests Amyloid β (Aβ) aggregation may cause the neurodegeneration associated with AD (63, 70, 132, 133).

Aβ aggregation plays a key role in the progression of Alzheimer's disease. Despite the importance of Aβ aggregation in disease etiology, our understanding of the determinants of aggregation is sparse and largely derived from *in vitro* studies. To overcome these gaps in knowledge, I use deep mutational scanning to measure the aggregation propensities of tens of thousands of Aβ variants. Recently, de Groot *et al.* (86) developed a yeast-based system to extricate toxicity from aggregation and thus offers a way to investigate the effects of protein sequence on aggregation propensity alone. In this system, Aβ is cytoplasmically localized to eliminate its aggregation-associated toxicity (71). To link Aβ aggregation to yeast growth, Aβ is fused to an essential protein, dihydrofolate reductase (DHFR) via a short peptide linker. The endogenously expressed concentration of DHF1 (yeast DHFR), is competitively inhibited by methotrexate. Furthermore, DHFR's activity is dependent upon Aβ solubility, and thus yeast with soluble Aβ variants rapidly grow in culture, whereas aggregating Aβ variants yield slow yeast growth.

High-throughput DNA sequencing tracked the frequency of each Aβ variant during the selection and enabled us to assign an aggregation propensity score to each variant in the library. Our study provides the first large-scale, in vivo mutational dataset of Aβ, which will illuminate the physicochemical properties of amino acids that negate, promote or do not effect Aβ aggregation. This data source has the potential to a novel way to identify correct models of aggregate structure.

4.2 RESULTS

4.2.1 *Multiplexing an assay for protein aggregation*

To begin, I verified that the Aβ-DHFR selection assay could be used to select for non-aggregating variants of Aβ in yeast (**Figure 4.1A**). Our ability to accurately measure the effects of Aβ mutations on aggregation propensity hinged upon the inhibition of DHFR with methotrexate. To identify the dose of methotrexate needed to inhibit endogenously expressed DHF1 concentrations, I performed a dose-response curve via yeast growth experiments. I find 80μM methotrexate to maximally distinguish yeast with aggregating (Aβ) and nonaggregating (Aβ-F19D) variants of Aβ-DHFR (**Figure 4.1B; Figure S4.1**).

Next, I performed a co-culture with Aβ-DHFR and Aβ(F19D)-DHFR to verify that our assay yields differential growth rates for aggregating and nonaggregating variants. In this experiment, co-culture samples were collected at 12 h intervals, plasmids were extracted and then prepped for Sanger sequencing to quantitate the proportions of Aβ variants in the culture. Indeed, I find that yeast expressing the nonaggregating variant out-grows wild-type Aβ (**Figure 4.1C**). Aggregation patterns, e.g. punctate vs. nonaggregating, for known aggregating and non-aggregating Aβ

variants were verified with fluorescence microscopy on Aβ-GFP variants. As expected, yeast expressing Aβ-GFP yield punctate aggregates after 24h induction, while yeast with Aβ(F19D)-GFP show diffuse fluorescence across the cell (**Figure 4.1D**). Thus, our assay preferentially selects for yeast with variants of Aβ that disrupt protein aggregation.

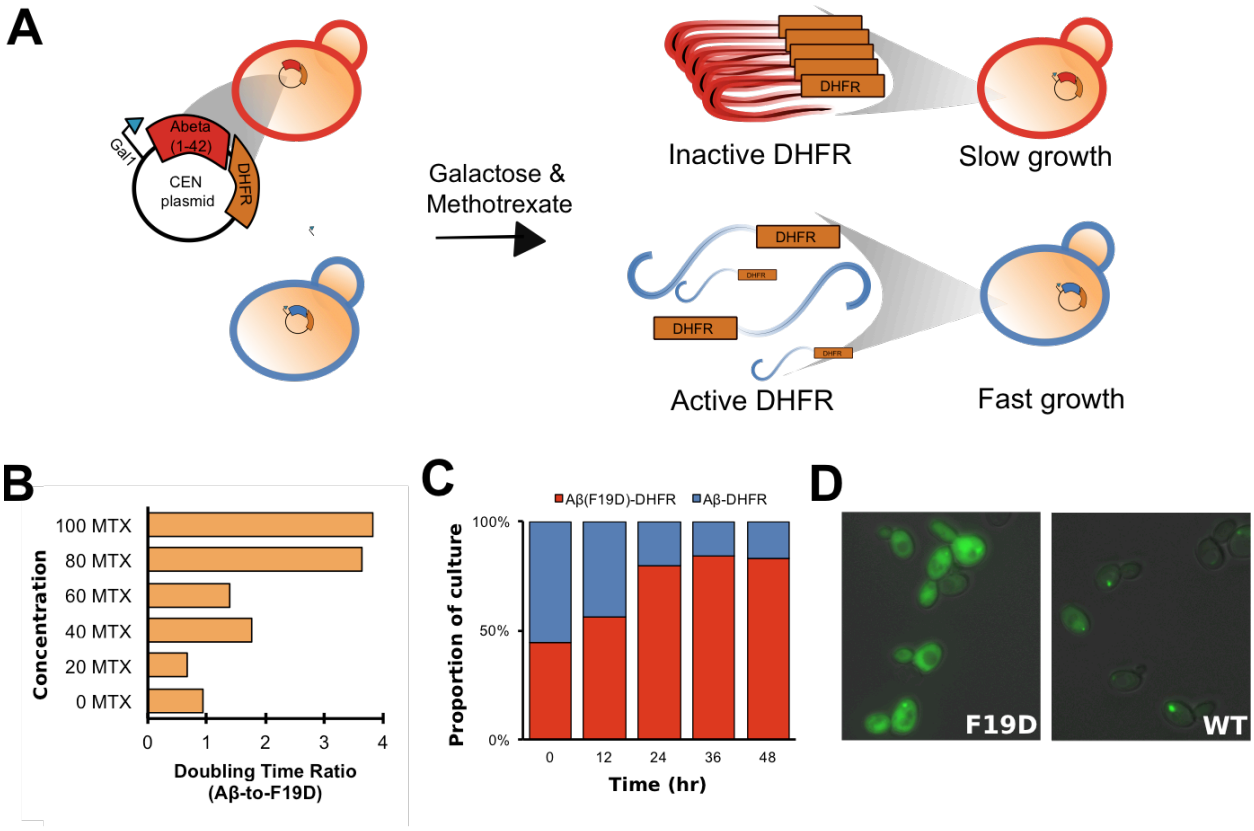


Figure 4-1. Yeast-based aggregation assay selects for non-aggregating variants of Aβ. An overview of the selection assay is shown in panel (A). Yeast are transformed with CEN plasmids that express Aβ variants fused to dihydrofolate reductase (DHFR) in the presence of galactose. Methotrexate competitively inhibits DHFR and in administered at a level that inhibits endogenously expressed DHFR concentrations (B). Non-aggregating variants of Aβ-DHFR rescue growth and increase in frequency faster than aggregating variants in co-culture (C). Aβ-GFP fusions show different aggregate patterns in aggregation- and nonaggregation-prone variants (D).

4.2.2 *Measuring aggregation propensity and replicability*

A library of Aβ variants was made using doped-oligo mutagenesis (See Methods). Aggregation propensities were measured for 15,238 (355 single amino acid) variants in the library (**Supplementary Table 4.1A-B**). As done by others (10, 15, 134), I used a statistical error cutoff guided by synonymous mutation effects across replicates to filter out potentially erroneous measures of protein aggregation. Here, I limit subsequent analyses to variants with standard errors <= 0.15. This cutoff yields an effective library size of 11,530 (330 single amino acid) variants. The aggregation propensity scores for filtered single amino acid mutations across triplicates are highly correlated (R=0.8-0.9) and suggest our assay reliably measures aggregation propensities (**Supplementary figure 4.1**).

4.2.3 *Effects of single amino acid mutations on aggregation propensity*

Aggregation behaviors of truncating and synonymous variants evince our assay's ability to measure aggregation (**Figure 4.2A**). I see exclusively low scores for mutations that terminate translation prior to DHFR because our assay is dependent upon DHFR activity rather than another phenotype like aggregation product toxicity. As expected, synonymous mutations also yield low scores because they presumably do not affect the native aggregation activity of Aβ. Additionally, I performed experiments to test estimate genetic drift in variant frequencies and the toxicity of Aβ variants. In these experiments, without Aβ expression (no galactose, no methotrexate) and without selection pressure (no methotrexate), respectively, I observe little to no effect of Aβ variants on yeast growth (**Figure S4.2-4**).

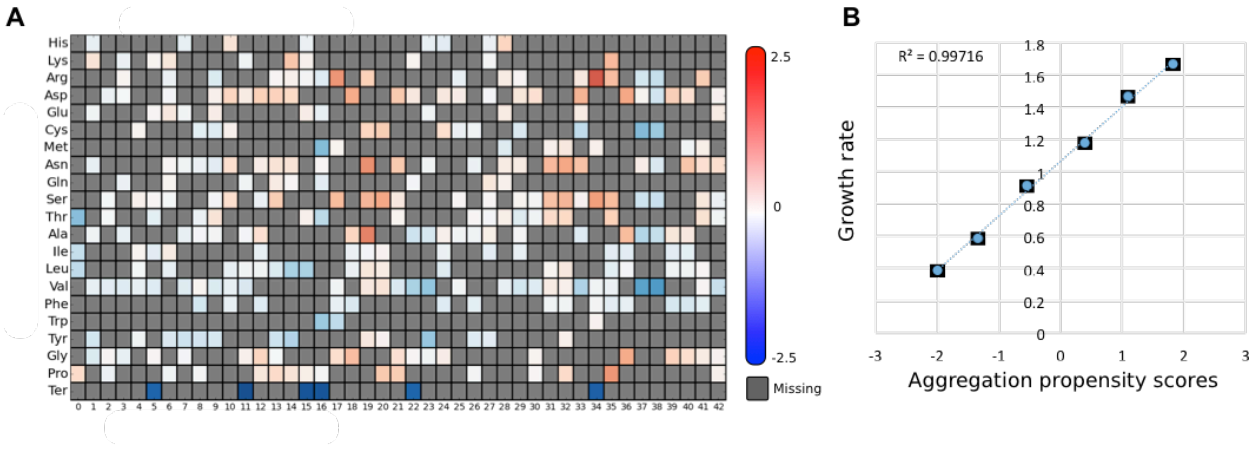


Figure 4-2. Effects of mutations on Aβ aggregation propensity. The aggregation propesnties of Aβ variants measured by deep mutation scanning are shown in a heatmap, where red and blue denote reduced and increased aggregation relative to wild-type Aβ (A). Six mutations that span the aggregation propensity were selected for validation via a low throughput assay where culture density was measured incrementally over a 24hr time period and then a regression line was fit to log phase growth to estimate variant growth rate (y-axis) (B).

To verify that our massively-parallel assay successfully measures variant effects, a handful of Aβ mutations that span aggregation propensity from highly soluble to aggregation-prone were subject to a low- throughput validation assay. In these experiments, the growth rate of yeast expressing individual Aβ variants were measured and compared to the aggregation propensity scores. Indeed, the low-throughput assay results are strongly correlated with the scores derived from sequencing read counts (**Figure 4.2B**). Thus, our assay can reliably measure the effects of Aβ variants on aggregation propensity.

To draw general conclusions about the amino acid determinants of protein aggregation, I investigated the quantitative effects of mutations along the Aβ sequence. First, I calculated the median effect of polar and nonpolar amino acid substitutions (**Figure 4.3A**). As expected, polar amino acids disrupt aggregation in sections of Aβ sequence harboring stretches of nonpolar amino acids. Yet, I find mutations to polar amino acids at some positions increase aggregation propensity. For example, polar amino acids increase aggregation propensity 36% at position 38,

which harbors a glycine residue. Congruently, I find mutations to nonpolar amino acids generally improve aggregation propensity. To determine whether this finding is a property of β-strands or potentially unique for amyloid-associated β-strands, I surveyed the effects of mutations in or flanking β-strands on protein function activity. β-strand and non-strand mutations come from 10 proteins with available deep mutational data. I find Aβ β-strands are more affected by polar

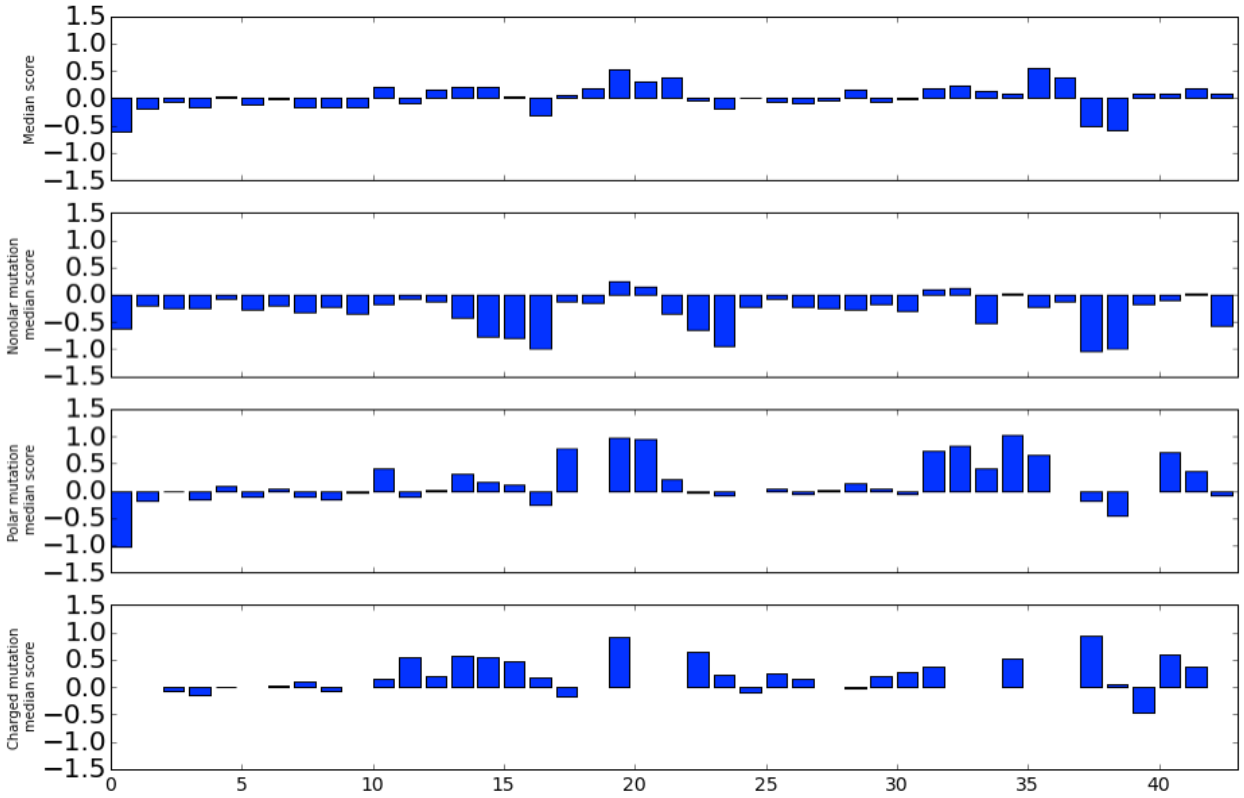


Figure 4-3 **Summary of mutational effects across Aβ sequence.** Barplots show the median of all, nonpolar (A,V,I,LF,M,Y,W), polar (S,T,N,Q) and charged (K,H,R,D,E) amino acids.

mutations than other β-strands, likely due to the fact aggregation is directly linked to β-strand potential, while other protein functions are less dependent upon it. (**Figure S4-5**).

To determine regions of Aβ sequence important for protein aggregation, I used agglomerative hierarchical clustering to mathematically distinguish between positions important and auxiliary for protein aggregation. As expected, positions in similar hydrophobicity classes cluster together.

For example, cluster 2 (in red) and 5 (in green) preponderantly harbor hydrophobic amino acid position, whereas clusters 1 and 4 largely contain polar positions. Interestingly, cluster 3 contains only 2 positions, which are both glycine and show increased aggregation propensities when mutated. Thus, hydrophobicity may be a major determinant in protein aggregation.

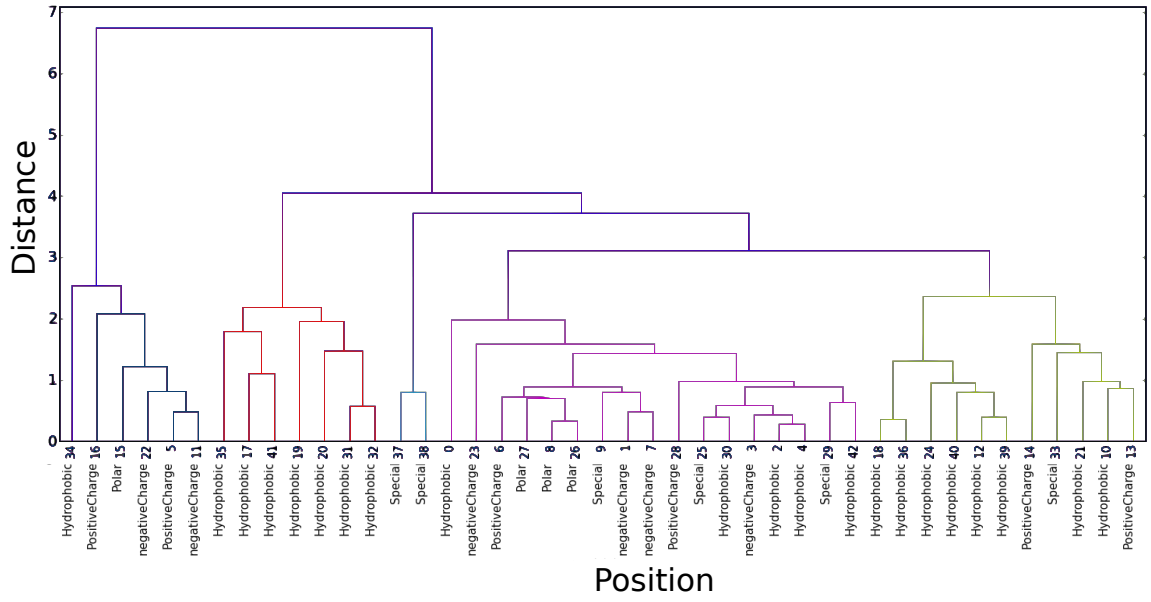


Figure 4-4 Aβ positions cluster by wild-type amino acid hydrophobicity. Hierarchical clustering using Ward's method was used on Aβ aggregation scores for each mutant amino acid at all positions. Dendrogram x-axis shows position and hydrophobicity of wild-type amino acid and y-axis shows the Ward's distance between clusters.

To draw general conclusions about the amino acid determinants of protein aggregation, the effects of mutations in the context of wild-type amino acids were investigated. The median effects of mutations for each amino acid type reveal that mutations to charged and polar amino acids generally increase aggregation propensities (**Figure 4-5A**). On the contrary, mutations at Aβ positions harboring hydrophobic residues are generally disruptive to aggregation. In a similar analysis, the median effects of mutations across mutant amino acid types were calculated. Here, I find a complimentary pattern, where mutations to hydrophobic amino acids increase aggregation

and polar or charged amino acids decrease it (**Figure 4-5B**).

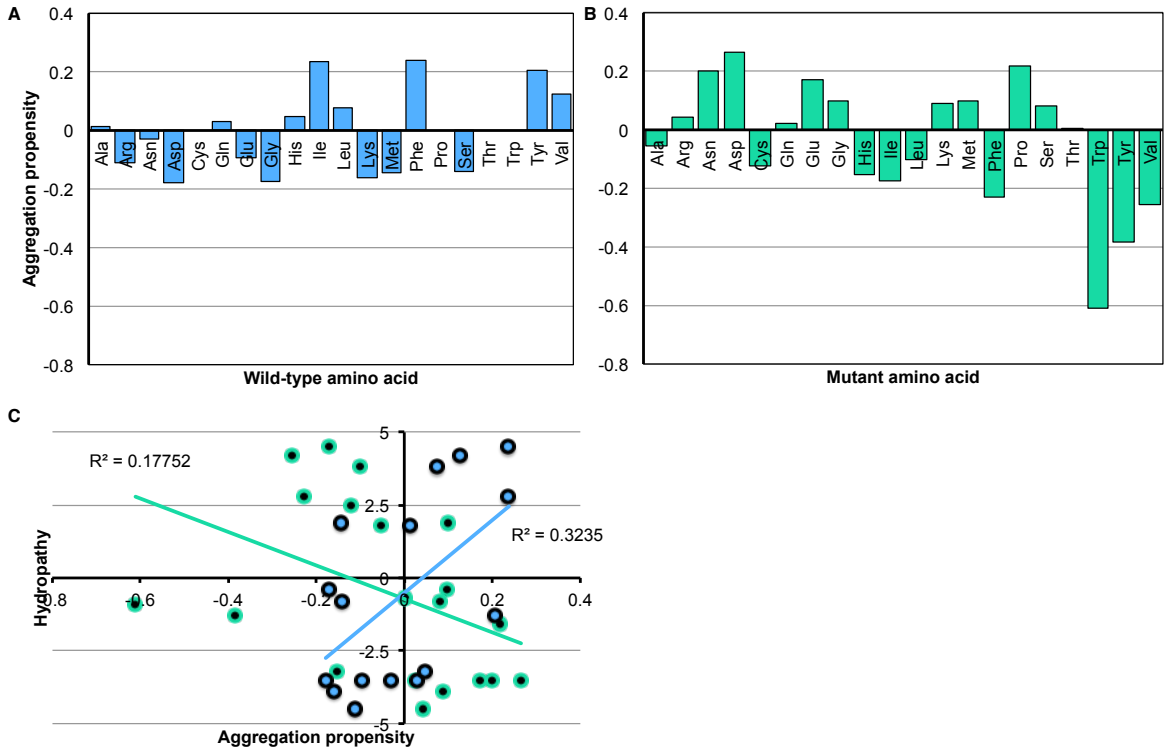


Figure 4-5. Hydrophobicity is a molecular determinant of protein aggregation. Median aggregation propensity scores for wild-type (**A**) and mutant (**B**) amino acids. The scatterplot shows the Pearsons R^2 correlation between hydropathy scores and the median aggregation propensity scores for wild-type and mutant amino acids (**C**).

To see whether hydrophobicity was linearly correlated with aggregation, hydropathy scores were compared to median aggregation propensity scores (**Figure 4-5C**). Indeed, some correlation between hydropathy and median aggregation scores exist, although the magnitude of correlation is not high (R^2 = 0.18 and 0.32). Thus, as suggested by our hierarchical clustering analysis, amino acid hydrophobicity is a determinant of protein aggregation.

The majority of Alzheimer's disease-associated mutations are not found in Aβ sequence, but instead in its precursor protein or the secretase proteins that process Aβ precursor protein.

However, several disease-associated mutations in Aβ sequence have been reported in Clinvar (135) . Additionally, Exac (113) reports additional mutations with unknown clinical significance that exist in humans. To investigate the relationship between aggregation and disease state, I collect the aforementioned mutations and compare them to our aggregation propensities (Figure 4-5A-B). I find no correlation between Exac mutant allele frequencies and aggregation propensities to exist. Of note, the mutation with highest allele frequency (Ala2Thr) maintains wild-type-like aggregation and five of seven pathogenic mutations also maintain aggregation propensity.

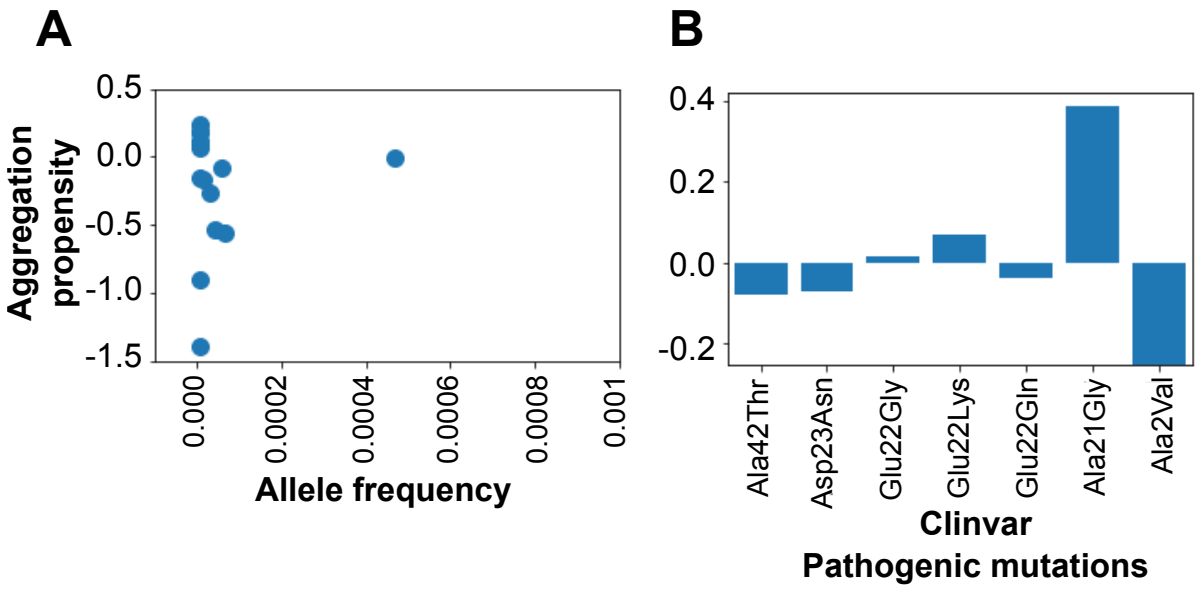


Figure 4-6. Aggregation propensity scores of human Aβ mutations. (A) A scatterplot compares Aβ mutant allele frequencies in humans to aggregation propensity scores. (B) A bar plot shows the aggregation propensities for seven pathogenic mutations from Clinvar.

4.3 METHODS

4.3.1 *Library construction*

The library was constructed *in vivo* as done by others (136). First, a forward primer containing a 5' homology region, an NNK codon, and a 3' extension region was designed for each codon in Aβ. The homology and extension regions were at least 15 nucleotides in length and had melting temperatures greater than 55C. Reverse primers were the reverse complement of the 5' homology region.

A unique PCR reaction was completed for each codon position. Each reaction contained 40ng template (p416GAL1-Aβ-DHFR) and 10 μM forward and reverse primers (IDT, custom oligos) in a total reaction volume of 30ul. The following cycling conditions were used: 95C 3min, 8x [98C 20sec, 60C 15sec, 72C 9min], 72C 9min. After PCR, 7.5μl of each product was run on a 1.5% agarose gel for 30min at 100V to check for a single product. The remaining 22.5μl aliquots of product were each digested for an hour at 37C with 0.6μl of DpnI (NEB, R0176S). After digestion, 4μl of each linear product was transformed into a 50μl of TOP10F Chemically Competent *E. coli* (ThermoFisher, C303003) according to manufacturer with the following modifications: the protocol was done in a 96 well plate, and cells were recovered in a total volume of 200μl SOC. After recovery, cells were transferred to a deep well plate with 1.6-1.8mL of ampicillin LB and shaken overnight. To get an estimate of colony count, 50μl of culture was plated on an LB + ampicillin agar plate. Deep well plates and agar plates were incubated at 37C overnight. After incubation, all 42 deep well plate cultures were combined and subject to a midiprep (Sigma, NA0200).

4.3.2 *Plasmids, yeast strains and growth conditions*

To create a galactose-regulated Aβ-DHFR expression system, I cloned the human Aβ coding sequence into the SpeI and HindIII sites of p416 (URA3, GAL1 promoter, CEN) (Mumberg, Müller, & Funk, 1994). Aβ-GFP variants were cloned using the same scheme. All Aβ variants were cloned into p416 and transformed in W303 strain (MATa/MATα {leu2-3,112 trp1-1 can1-100 ura3-1 ade2-1 his3-11,15} [phi+]). Cells were grown at 30°C in synthetic complete (SC) media lacking uracil and supplemented with 2% glucose. For aggregation experiments Aβ library expression was induced with 2% galactose in SC lacking uracil and adding the DHFR inhibitor, methotrexate (TCI America, M-1664), to a final concentration of 80 μM and 1mM sulfanilamide (Sigma, S-9251), unless otherwise indicated. A similar experiment was conducted without methotrexate the test the effects of Aβ expression on yeast growth and a another experiment was conducted without methotrexate or galactose induction to estimate genetic drift.

4.3.3 *Selection assay*

Transformed yeast were inoculated into 5mL or 300 mL of C-Ura, 2% glucose media, grown in a rotating/shaking, 30C incubator over night and then transferred to 5 mL or 300 mL 2% raffinose media to remove the glucose repression acting on the *gal1* promoter. After two hours in 2% raffinose, yeast were back-diluted to an OD of 0.01 into 5 mL or 300 mL 2% galactose to induce gene expression in the presence and absence of 80μM methotrexate, 1mM sulfanilamide. For 300 mL experiments, wherein yeast with aggregating and nonaggregating variants were grown in co-culture, two strains were inoculated at a 1:1. In 5mL experiments, yeast growth was measured using a spectrophotometer that detects 660 nm wavelengths throughout a 48h course. In

competition experiments, 10 OD units of yeast were collected from 300 mL cultures every 12h, spun down, concentrated and stored in -80C until experiment concluded. Frozen yeast were then thawed and then their plasmids were extracted using DNA clean and concentrator kit (Zymo Research). Extracted plasmids were prepped for sequencing and sequenced using Sanger Sequencing. I used the following equation to calculate doubling times from two time points: $(Log_{10}(OD_{T2}/OD_{T1})/ Log_{10}(2))/ \Delta T$, where OD represents the optical density at 600nm at a time point (T).

4.3.4 *Library preparation for high-throughput sequencing*

Plasmids were extracted from yeast with ZYMO Research Zymoprep Yeast Plasmid Miniprep 1 kit (Zymo Research, D-2001). Library fragments were amplified by 17 PCR cycles using primers specific to DNA sequences that flank Aβ-DHFR in p416, and sequenced by an Illumina NextSeq sequencer by pair-end reads.

4.3.5 *Variant effect analysis*

Functional scores are calculated by dividing the frequency of a particular variant after selection to the frequency of that variant before selection, which is then normalized by the rate of change for wild-type sequence and then log_2 transformed. Scores below 0 denote variants that are more aggregation-prone that wild-type, whereas scores above 0 indicate that a variant has low aggregation propensity.

4.4 DISCUSSION

I have laid the foundational analyses for identifying the molecular determinants of aggregation. By adapting a low-throughput, yeast-based assay to measure Aβ variant effects in high-throughput, I was able to quantitate the effects of ~350 single amino acid mutations on protein aggregation propensity. I used these mutational effects to create a sequence function map, which reveals information about positions and regions of Aβ sequence most important for protein aggregation. Some Aβ mutations are associated with disease, such as E22Q and D23N, which are associated with cerebral Aβ angiopathy (137, 138). Other studies report these mutations rapidly form fibrils, while our assay revealed only mild increase in aggregation relative to wild-type(139). This difference in results could be due to other studies working with Aβ 1-40, rather than 42, in *in vitro* settings or high assay sensitivity. Nonetheless my sequence-function map can be used as a lookup table for previously unobserved Aβ mutations in humans.

The molecular effects of mutations of Aβ aggregation will likely vary at positions participating in β-strand structure, turns and unstructured regions, and while many models of aggregate structure exist, concordance in secondary structures among models is low. Thus, I aimed to characterize regions of Aβ sequence with mutational effect patterns. Because polar and nonpolar amino acid mutations are expected to maintain and disrupt β-strands and consequently protein aggregation, I will use these sets of amino acid types to probe Aβ sequence for β-strands. To verify that polar and nonpolar amino acid mutations effect Aβ aggregation propensity as expected, I surveyed the effects of all, polar, nonpolar and charged amino acids across sequence. Next, I showed that Aβ positions cluster according to the hydrophobicities of the wild-type amino acids. These results suggest that using polar and nonpolar amino acids to probe for β-

strands may be a viable approach to revealing Aβ secondary structure. This proposed future analysis will offer a unique opportunity determine which models of Aβ aggregate structure best support the aggregate structures forming in an *in vivo* system.

In an analysis on Clinvar and Exac mutations, I've showed that aggregation propensity scores and disease-state are not directly correlated. This correlation could exist for two reasons. First, many mutations associated with Alzheimer's disease or other Aβ-opathies, effect the cleavage of Aβ from the Aβ precursor protein (APP). Because our assay is agnostic to APP processing, such effects will be missed and is a limitation of the experimental setup. Second, Alzheimer's is a very complex disease and the Aβ aggregation I measure may not be the underlying or only cause of the disease. For instance, most evidence shows Aβ oligomers to be the neurotoxic species, however, I currently do not know the limits of our assay to detect oligomerization, fibril formation and generalized aggregation product. Thus, additional assays focused on APP processing should be applied to fully understand the relationship between Aβ mutation and disease.

Chapter 5. OUTSTANDING QUESTIONS AND FUTURE LARGE-SCALE MUTAGENESIS ANALYSES

5.1 COMPONENTIZING PREDICTIONS

Componentizing is a software concept that suggests systems should be developed as an assembly of modular parts (140). For instance, instead of universal software, software is built of components that can be deployed independently. This concept gained popularity as software engineering transitioned from marginal to central across fields. In Chapter 3, I introduced Envision, a predictor for the molecular effect of mutations on protein function. Upon Envision's conception, only a handful of large-scale mutagenesis datasets were available for model training. Thus, it was not possible to train a variant effect model with the ability to predict specific protein functions (*e.g.* binding decreased by 10%). However, since then, over 70 deep mutational scanning datasets have been published and this rapidly amassing pool of data provides new opportunities to train models to predict high-resolution.

To effectively train high-resolution predictors, I foresee applying component-based thinking to model training. First, large-scale mutagenesis datasets will be categorized by assay type (e.g. binding, stability, enzymatic activity). Second, unique predictors need to be trained for each assay type. Third, stacked generalization, which is a machine learning technique, can be used to combine models. Stacked generalization provides a way to combine predictors to yield an ultimate prediction for mutational effect, but also retain high-resolution predictions.

5.2 Predicting Pathogenicity

As observed in Chapter 3, Envision does not predict pathogenicity well. This is likely due to the fact Envision was not trained to predict disease-associated mutations, but rather the quantitative effects of mutations on protein function. A number of approaches could lead to an accurate pathogenicity predictor. The first and easiest approach would be to combine Envision with a state-of-the-art pathogenicity predictor. Naïve Bayes and logistic regression are two available methods for learning the thresholds from two predictors that best discern pathogenic from benign mutations.

The relationship between mutational effect magnitude and pathogenicity is not known for every protein. For instance, we do not know whether 50% or 20% decrease in protein activity of a gene is enough to incur disease. Here, I propose a second approach that aims to learn the relationships between quantitative protein function and pathogenicity. To do this, one could perform several deep mutational scans on proteins highly mutated in the human population and associated with disease. It should then be possible to identify how disruptive a mutation needs to be to cause disease for each protein included in the study. Further analyses can reveal whether general patterns govern the level of protein activity associated with disease and health. If general patterns do exist, a machine-learning algorithm may learn how to draw cutoffs in large-scale mutagenesis datasets to separate pathogenic from non-pathogenic mutations. If general patterns do not exist, but instead each protein function has a unique relationship with disease, each protein will need to be studied individually.

5.3 USING ROSETTA TO WINNOW IN VIVO MODELS OF AGGREGATE STRUCTURE

Rosetta is a computational suite designed for modeling and analysis of protein structures. One feature of Rosetta is its score function (141). Score functions in Rosetta are weighted sums of energy terms, which represent electrostatics and van der Waals' interaction physical forces and statistical terms like the probability of finding the torsion angles in Ramachandran space. In Rosetta, low scores indicate more likely native structures than higher scores. As mentioned in Chapter 4, many models of Aβ aggregate structure exist. Most models are based on *in vitro* derived aggregates and solid-state NMR evidence, which paints a far-from-complete picture of aggregate structure. Thus, it is not surprising that available models are highly diverse in tertiary and secondary structure. I am interested in using my large-scale mutagenesis data to winnow down the set of candidate *in vivo* aggregate models. To do this, I propose to compare my Aβ aggregation scores to Rosetta energy scores for each available aggregate model. I expect for the model most consistent with my data, to show the highest correlation between energy scores and aggregation propensities.

The protein structure field has long known that evolutionary covariance or epistasis can be used to identify contacting positions within or between proteins (142-145). Recently, a study revealed that epistatic positions found in deep mutational scans of doubly and singly mutated variants can also be used to identify pairs of positions that are proximal in a protein structure (146). My deep mutational scan of Aβ contains ~33% of all possible double mutations and thus provides and opportunity to identify pairs of candidate positions that may be in contact in Aβ aggregate structure. Aβ models could be winnowed down by analyzing the candidate contact positions

within the context of each model of aggregate structure. The model that contains the most candidate pairs of positions is the model that best reflects my aggregation propensity scores.

BIBLIOGRAPHY

1. Winter,G., Fersht,A.R., Wilkinson,A.J., Zoller,M. and Smith,M. (1982) Redesigning enzyme structure by site-directed mutagenesis: tyrosyl tRNA synthetase and ATP binding. *Nature*, **299**, 756–758.

2. Hutchison,C.A., Phillips,S., Edgell,M.H., Gillam,S., Jahnke,P. and Smith,M. (1978) Mutagenesis at a specific position in a DNA sequence. *J. Biol. Chem.*, **253**, 6551–6560.

3. Cunningham,B.C. and Wells,J.A. (1989) High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science*, **244**, 1081–1085.

4. Fowler,D.M. and Fields,S. (2014) Deep mutational scanning: a new style of protein science. *Nat. Methods*, **11**, 801–807.

5. Fowler,D.M., Araya,C.L., Fleishman,S.J., Kellogg,E.H., Stephany,J.J., Baker,D. and Fields,S. (2010) High-resolution mapping of protein sequence-function relationships. *Nat. Methods*, **7**, 741–746.

6. High-resolution mapping of protein sequence-function relationships (2010) High-resolution mapping of protein sequence-function relationships. **7**, 741–746.

7. Knight,R. and Yarus,M. (2003) Analyzing partially randomized nucleic acid pools: straight dope on doping. *Nucl. Acids Res.*, **31**, e30.

8. Jain,P.C. and Varadarajan,R. (2014) A rapid, efficient, and economical inverse polymerase chain reaction-based method for generating a site saturation mutant library. *Anal. Biochem.*, **449**, 90–98.

9. Mishra,P., Flynn,J.M., Starr,T.N. and Bolon,D.N.A. (2016) Systematic Mutant Analyses Elucidate General and Client-Specific Aspects of Hsp90 Function. *Cell Reports*, **15**, 588–598.

10. Melamed,D., Young,D.L., Gamble,C.E., Miller,C.R. and Fields,S. (2013) Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA*, **19**, 1537–1551.

11. Starita,L.M., Pruneda,J.N., Lo,R.S., Fowler,D.M., Kim,H.J., Hiatt,J.B., Shendure,J., Brzovic,P.S., Fields,S. and Klevit,R.E. (2013) Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc. Natl. Acad. Sci.*, **110**, 1263–1272.

12. Roscoe,B.P., Thayer,K.M., Zeldovich,K.B., Fushman,D. and Bolon,D.N.A. (2013) Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J. Mol. Biol.*, **425**, 1363–1377.

13. Firnberg,E., Labonte,J.W., Gray,J.J. and Ostermeier,M. (2014) A comprehensive, high-resolution map of a gene's fitness landscape. *Mol. Biol. Evol.*, **31**, 1581–1592.

14. Melnikov,A., Rogov,P., Wang,L., Gnirke,A. and Mikkelsen,T.S. (2014) Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucl. Acids Res.*, **42**, 1–8.

15. Starita,L.M., Young,D.L., Islam,M., Kitzman,J.O., Gullingsrud,J., Hause,R.J., Fowler,D.M., Parvin,J.D., Shendure,J. and Fields,S. (2015) Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics*, **200**, 413–422.

16. Deng,C.X. and Brodie,S.G. (2000) Roles of BRCA1 and its interacting proteins. *Bioessays*, **22**, 728–737.

17. Sarkisyan,K.S., Bolotin,D.A., Meer,M.V., Usmanova,D.R., Mishin,A.S., Sharonov,G.V., Ivankov,D.N., Bozhanova,N.G., Baranov,M.S., Soylemez,O., *et al.* (2016) Local fitness landscape of the green fluorescent protein. *Nature*, **533**, 397–401.

18. Bhagavatula,G., Rich,M.S., Young,D.L., Marin,M. and Fields,S. (2017) A Massively Parallel Fluorescence Assay to Characterize the Effects of Synonymous Mutations on TP53 Expression. *Mol. Cancer Res.*, **15**, 1301–1307.

19. Matreyek,K.A., Stephany,J.J. and Fowler,D.M. (2017) A platform for functional assessment of large variant libraries in mammalian cells. *Nucl. Acids Res.*, **45**, e102–e102.

20. Fowler,D.M. and Fields,S. (2014) Deep mutational scanning: a new style of protein science. *Nat. Methods*, **11**, 801–807.

21. Fowler,D.M., Stephany,J.J. and Fields,S. (2014) Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat. Protoc.*, **9**, 2267–2284.

22. Loman,N.J., Misra,R.V., Dallman,T.J., Constantinidou,C., Gharbia,S.E., Wain,J. and Pallen,M.J. (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.*, **30**, 434–439.

23. Goodwin,S., McPherson,J.D. and McCombie,W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, **17**, 333–351.

24. Bahassi,E.M. and Stambrook,P.J. (2014) Next-generation sequencing technologies: breaking the sound barrier of human genetics. *Mutagenesis*, **29**, 303–310.

25. Ross,M.G., Russ,C., Costello,M., Hollinger,A., Lennon,N.J., Hegarty,R., Nusbaum,C. and Jaffe,D.B. (2013) Characterizing and measuring bias in sequence data. *Genome Biol.*, **14**, R51.

26. Magrane,M. and UniProt Consortium (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database*, **2011**, 1–13.

27. Hiatt,J.B., Patwardhan,R.P., Turner,E.H., Lee,C. and Shendure,J. (2010) Parallel, tag-directed assembly of locally derived short sequence reads. *Nat. Methods*, **7**, 119–122.

28. Quail,M.A., Smith,M., Coupland,P., Otto,T.D., Harris,S.R., Connor,T.R., Bertoni,A., Swerdlow,H.P. and Gu,Y. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.

29. Rubin,A.F., Gelman,H., Lucas,N., Bajjalieh,S.M., Papenfuss,A.T., Speed,T.P. and Fowler,D.M. (2017) A statistical framework for analyzing deep mutational scanning data. *Genome Biol.*, **18**, 150.

30. Gasperini,M., Starita,L. and Shendure,J. (2016) The power of multiplexed functional analysis of genetic variants. *Nat. Protoc.*, **11**, 1782–1787.

31. Collins,F.S., Brooks,L.D. and Chakravarti,A. (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.*, **8**, 1229–1231.

32. Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.

33. Ramensky,V., Bork,P. and Sunyaev,S. (2002) Human non-synonymous SNPs: server and survey. *Nucl. Acids Res.*, **30**, 3894–3900.

34. Sim,N.-L., Kumar,P., Hu,J., Henikoff,S., Schneider,G. and Ng,P.C. (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucl. Acids Res.*, **40**, 452–457.

35. Kumar,P., Henikoff,S. and Ng,P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.

36. McLaren,W., Pritchard,B., Rios,D., Chen,Y., Flicek,P. and Cunningham,F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, **26**, 2069–2070.

37. Capriotti,E., Fariselli,P., Rossi,I. and Casadio,R. (2008) A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics*, **9**, S6.

38. Katsonis,P. and Lichtarge,O. (2014) A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. *Genome Res.*, **24**, 2050–2058.

39. Ferrer Costa,C., Orozco,M. and la Cruz,de,X. (2004) Sequence□based prediction of pathological mutations. *Proteins: Structure, Function, and Bioinformatics*, **57**, 811–819.

40. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.

41. Cingolani,P., Platts,A., Le Lily Wang, Coon,M., Nguyen,T., Wang,L., Land,S.J., Lu,X. and Ruden,D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *http://dx.doi.org/10.4161/fly.19695*, **6**, 80–92.

42. Hecht,M., Bromberg,Y. and Rost,B. (2015) Better prediction of functional effects for sequence variants. *BMC Genomics*, **16**, 1–12.

43. Pires,D.E.V., Ascher,D.B. and Blundell,T.L. (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucl. Acids Res.*, **42**, gku411–W319.

44. Kumar,S., Sanderford,M., Gray,V.E., Ye,J. and Liu,L. (2012) Evolutionary diagnosis method for variants in personal exomes. *Nat. Methods*, **9**, 855–856.

45. Jagadeesh,K.A., Wenger,A.M., Berger,M.J., Guturu,H., Stenson,P.D., Cooper,D.N., Bernstein,J.A. and Bejerano,G. (2016) M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.*, **48**, 1581–1586.

46. Thusberg,J., Olatubosun,A. and Vihinen,M. (2011) Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.*, **32**, 358–368.

47. Pires,D.E.V., Ascher,D.B. and Blundell,T.L. (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, **30**, 335–342.

48. Tang,H. and Thomas,P.D. (2016) Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. *Genetics*, **203**, 635–647.

49. Hopf,T.A., Ingraham,J.B., Poelwijk,F.J., Schärfe,C.P.I., Springer,M., Sander,C. and Marks,D.S. (2017) Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.*, **35**, 128–135.

50. Mahmood,K., Jung,C.-H., Philip,G., Georgeson,P., Chung,J., Pope,B.J. and Park,D.J. (2017) Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. *Human Genomics*, **11**, 10.

51. Friedman,J.H. (2002) Stochastic gradient boosting. *Comput. Stat. Data Anal.*, **38**, 367–378.

52. Freund,Y. (2009) A more robust boosting algorithm. *arXiv*, **stat.ML**.

53. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

54. Gray,V.E., Kukurba,K.R. and Kumar,S. (2012) Performance of computational tools in evaluating the functional impact of laboratory-induced amino acid mutations. *Bioinformatics*, **28**, 2093–2096.

55. Liu,L. and Kumar,S. (2013) Evolutionary balancing is critical for correctly forecasting disease-associated amino acid variants. *Mol. Biol. Evol.*, **30**, 1252–1257.

56. Sawaya,M.R., Sambashivan,S., Nelson,R., Ivanova,M.I., Sievers,S.A., Apostol,M.I., Thompson,M.J., Balbirnie,M., Wiltzius,J.J.W., McFarlane,H.T., *et al.* (2007) Atomic

structures of amyloid cross-beta spines reveal varied steric zippers. *Nature*, **447**, 453–457.

57. Eanes,E.D. and Glenner,G.G. (1968) X-ray diffraction studies on amyloid filaments. *J. Histochem. Cytochem.*, **16**, 673–677.

58. Tycko,R. and Wickner,R.B. (2013) Molecular structures of amyloid and prion fibrils: consensus versus controversy. *Acc. Chem. Res.*, **46**, 1487–1496.

59. Nelson,R., Sawaya,M.R., Balbirnie,M., Madsen,A.Ø., Riekel,C., Grothe,R. and Eisenberg,D. (2005) Structure of the cross-|[beta]| spine of amyloid-like fibrils. *Nature*, **435**, 773–778.

60. Glenner,G.G. and Wong,C.W. (1984) Alzheimer's disease: initial report of the purification and characterization of a novel cerebrovascular amyloid protein. *Biochem. Biophys. Res. Commun.*, **120**, 885–890.

61. Alzheimer,A., Stelzmann,R.A., Schnitzlein,H.N. and Murtagh,F.R. (1995) An English translation of Alzheimer's 1907 paper, "Uber eine eigenartige Erkankung der Hirnrinde". Wiley Subscription Services, Inc., A Wiley Company.

62. Hebert,L.E., Weuve,J., Scherr,P.A. and Evans,D.A. (2013) Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. *Neurology*, **80**, 1778–1783.

63. Lambert,M.A., Bickel,H., Prince,M., Fratiglioni,L., Strauss,Von,E., Frydecka,D., Kiejna,A., Georges,J. and Reynish,E.L. (2014) Estimating the burden of early onset dementia; systematic review of disease prevalence. *Eur. J. Neurol.*, **21**, 563–569.

64. Lee,J.-H., Jeong,S.-K., Kim,B.C., Park,K.W. and Dash,A. (2015) Donepezil across the spectrum of Alzheimer's disease: dose optimization and clinical relevance. *Acta Neurol. Scand.*, **131**, 259–267.

65. Jarrett,J.T., Berger,E.P. and Lansbury,P.T. (1993) The carboxy terminus of the beta amyloid protein is critical for the seeding of amyloid formation: implications for the pathogenesis of Alzheimer's disease. *Biochemistry*, **32**, 4693–4697.

66. Hartmann,T., Bieger,S.C., Brühl,B., Tienari,P.J., Ida,N., Allsop,D., Roberts,G.W., Masters,C.L., Dotti,C.G., Unsicker,K., *et al.* (1997) Distinct sites of intracellular production for Alzheimer's disease A beta40/42 amyloid peptides. *Nat. Med.*, **3**, 1016–1020.

67. Gunawardena,S. and Goldstein,L.S. (2001) Disruption of axonal transport and neuronal viability by amyloid precursor protein mutations in Drosophila. *Neuron*, **32**, 389–401.

68. Shoji,M., Golde,T.E., Ghiso,J., Cheung,T.T., Estus,S., Shaffer,L.M., Cai,X.D., McKay,D.M., Tintner,R. and Frangione,B. (1992) Production of the Alzheimer amyloid beta protein by normal proteolytic processing. *Science*, **258**, 126–129.

69. Tanzi,R.E., Gusella,J.F., Watkins,P.C., Bruns,G.A., St George-Hyslop,P., Van Keuren,M.L., Patterson,D., Pagan,S., Kurnit,D.M. and Neve,R.L. (1987) Amyloid beta protein gene: cDNA, mRNA distribution, and genetic linkage near the Alzheimer locus. *Science*, **235**,

880–884.

70. Haass,C., Hung,A.Y., Selkoe,D.J. and Teplow,D.B. (1994) Mutations associated with a locus for familial Alzheimer's disease result in alternative processing of amyloid beta-protein precursor. *J. Biol. Chem.*, **269**, 17741–17748.

71. Vassar,R., Bennett,B.D., Babu-Khan,S., Kahn,S., Mendiaz,E.A., Denis,P., Teplow,D.B., Ross,S., Amarante,P., Loeloff,R., *et al.* (1999) Beta-secretase cleavage of Alzheimer's amyloid precursor protein by the transmembrane aspartic protease BACE. *Science*, **286**, 735–741.

72. Dahlgren,K.N., Manelli,A.M., Stine,W.B., Baker,L.K., Krafft,G.A. and LaDu,M.J. (2002) Oligomeric and fibrillar species of amyloid-beta peptides differentially affect neuronal viability. *J. Biol. Chem.*, **277**, 32046–32053.

73. Iwatsubo,T., Odaka,A., Suzuki,N., Mizusawa,H., Nukina,N. and Ihara,Y. (1994) Visualization of A beta 42(43) and A beta 40 in senile plaques with end-specific A beta monoclonals: evidence that an initially deposited species is A beta 42(43). *Neuron*, **13**, 45–53.

74. Kane,M.D., Lipinski,W.J., Callahan,M.J., Bian,F., Durham,R.A., Schwarz,R.D., Roher,A.E. and Walker,L.C. (2000) Evidence for seeding of beta -amyloid by intracerebral infusion of Alzheimer brain extracts in beta -amyloid precursor protein-transgenic mice. *J. Neurosci.*, **20**, 3606–3611.

75. Wirths,O., Multhaup,G., Czech,C., Blanchard,V., Moussaoui,S., Tremp,G., Pradier,L., Beyreuther,K. and Bayer,T.A. (2001) Intraneuronal Abeta accumulation precedes plaque formation in beta-amyloid precursor protein and presenilin-1 double-transgenic mice. *Neurosci. Lett.*, **306**, 116–120.

76. D'Andrea,M.R., Nagele,R.G., Wang,H.Y., Peterson,P.A. and Lee,D.H. (2001) Evidence that neurones accumulating amyloid can undergo lysis to form amyloid plaques in Alzheimer's disease. *Histopathology*, **38**, 120–134.

77. Walsh,D.M., Tseng,B.P., Rydel,R.E., Podlisny,M.B. and Selkoe,D.J. (2000) The oligomerization of amyloid beta-protein begins intracellularly in cells derived from human brain. *Biochemistry*, **39**, 10831–10839.

78. Benilova,I., Karran,E. and De Strooper,B. (2012) The toxic Aβ oligomer and Alzheimer's disease: an emperor in need of clothes. *Nat. Neurosci.*, **15**, 349–357.

79. Pike,C.J., Walencewicz,A.J., Glabe,C.G. and Cotman,C.W. (1991) In vitro aging of beta-amyloid protein causes peptide aggregation and neurotoxicity. *Brain Res.*, **563**, 311–314.

80. McLean,C.A., Cherny,R.A., Fraser,F.W., Fuller,S.J., Smith,M.J., Beyreuther,K., Bush,A.I. and Masters,C.L. (1999) Soluble pool of Abeta amyloid as a determinant of severity of neurodegeneration in Alzheimer's disease. *Ann. Neurol.*, **46**, 860–866.

81. Antzutkin,O.N., Leapman,R.D., Balbach,J.J. and Tycko,R. (2002) Supramolecular structural constraints on Alzheimer's beta-amyloid fibrils from electron microscopy and solid-state nuclear magnetic resonance. *Biochemistry*, **41**, 15436–15450.

82. Lührs,T., Ritter,C., Adrian,M., Riek-Loher,D., Bohrmann,B., Döbeli,H., Schubert,D. and Riek,R. (2005) 3D structure of Alzheimer's amyloid-beta(1-42) fibrils. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 17342–17347.

83. Petkova,A.T., Ishii,Y., Balbach,J.J., Antzutkin,O.N., Leapman,R.D., Delaglio,F. and Tycko,R. (2002) A structural model for Alzheimer's beta -amyloid fibrils based on experimental constraints from solid state NMR. *Proc. Natl. Acad. Sci.*, **99**, 16742–16747.

84. Treusch,S., Hamamichi,S., Goodman,J.L., Matlack,K.E.S., Chung,C.Y., Baru,V., Shulman,J.M., Parrado,A., Bevis,B.J., Valastyan,J.S., *et al.* (2011) Functional links between Aβ toxicity, endocytic trafficking, and Alzheimer's disease risk factors in yeast. *Science*, **334**, 1241–1245.

85. Chakrabortee,S., Byers,J.S., Jones,S., Garcia,D.M., Bhullar,B., Chang,A., She,R., Lee,L., Fremin,B., Lindquist,S., *et al.* (2016) Intrinsically Disordered Proteins Drive Emergence and Inheritance of Biological Traits. *Cell*, **167**, 369–381.e12.

86. Morell,M., de Groot,N.S., Vendrell,J., Avilés,F.X. and Ventura,S. (2011) Linking amyloid protein aggregation and yeast survival. *Mol Biosyst*, **7**, 1121–1128.

87. Nanevicz,T., Ishii,M., Wang,L., Chen,M., Chen,J., Turck,C.W., Cohen,F.E. and Coughlin,S.R. (1995) Mechanisms of thrombin receptor agonist specificity. Chimeric receptors and complementary mutations identify an agonist recognition site. *J. Biol. Chem.*, **270**, 21619–21625.

88. Kanaya,E., Kanaya,S. and Kikuchi,M. (1990) Introduction of a non-native disulfide bridge to human lysozyme by cysteine scanning mutagenesis. *Biochem. Biophys. Res. Commun.*, **173**, 1194–1199.

89. Valbuena,J.J., Vera,R., García,J., Puentes,A., Curtidor,H., Ocampo,M., Urquiza,M., Rivera,Z., Guzmán,F., Torres,E., *et al.* (2003) Plasmodium falciparum normocyte binding protein (PfNBP-1) peptides bind specifically to human erythrocytes. *Peptides*, **24**, 1007–1014.

90. Woods,A.C., Guillemette,J.G., Parrish,J.C., Smith,M. and Wallace,C.J. (1996) Synergy in protein engineering. Mutagenic manipulation of protein structure to simplify semisynthesis. *J. Biol. Chem.*, **271**, 32008–32015.

91. Borngräber,S., Browner,M., Gillmor,S., Gerth,C., Anton,M., Fletterick,R. and Kühn,H. (1999) Shape and specificity in mammalian 15-lipoxygenase active site. The functional interplay of sequence determinants for the reaction specificity. *J. Biol. Chem.*, **274**, 37345–37350.

92. Vandemeulebroucke,A., De Vos,S., Van Holsbeke,E., Steyaert,J. and Versées,W. (2008) A

flexible loop as a functional element in the catalytic mechanism of nucleoside hydrolase from Trypanosoma vivax. *J. Biol. Chem.*, **283**, 22272–22282.

93. Zhang,L., Wang,L., Kao,Y.-T., Qiu,W., Yang,Y., Okobiah,O. and Zhong,D. (2007) Mapping hydration dynamics around a protein surface. *Proc. Natl. Acad. Sci.*, **104**, 18461–18466.

94. Bromberg,Y. and Rost,B. (2008) Comprehensive in silico mutagenesis highlights functionally important residues in proteins. *Bioinformatics*, **24**, 207–212.

95. Bromberg,Y., Overton,J., Vaisse,C., Leibel,R.L. and Rost,B. (2009) In silico mutagenesis: a case study of the melanocortin 4 receptor. *FASEB J.*, **23**, 3059–3069.

96. Xiao,Y., Wigneshweraraj,S.R., Weinzierl,R., Wang,Y.-P. and Buck,M. (2009) Construction and functional analyses of a comprehensive sigma54 site-directed mutant library using alanine-cysteine mutagenesis. *Nucl. Acids Res.*, **37**, 4482–4497.

97. Bromberg,Y. and Rost,B. (2008) Comprehensive in silico mutagenesis highlights functionally important residues in proteins. *Bioinformatics*, **24**, 207–212.

98. Dayhoff,M.O. (1978) Atlas of Protein Sequence and Structure.

99. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.*, **89**, 10915–10919.

100. Grantham,R. (1974) Amino acid difference formula to help explain protein evolution. *Science*, **185**, 862–864.

101. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

102. Costantini,S., Colonna,G. and Facchiano,A.M. (2006) Amino acid propensities for secondary structures are influenced by the protein structural class. *Biochem. Biophys. Res. Commun.*, **342**, 441–451.

103. Chen,H. and Zhou,H.-X. (2005) Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucl. Acids Res.*, **33**, 3193–3199.

104. Doyle,D.A., Lee,A., Lewis,J., Kim,E., Sheng,M. and MacKinnon,R. (1996) Crystal Structures of a Complexed and Peptide-Free Membrane Protein–Binding Domain: Molecular Basis of Peptide Recognition by PDZ. *Cell*, **85**, 1067–1076.

105. Marmorstein,R. and Carey,M. (1992) DNA recognition by GAL4: structure of a protein-DNA complex. *Nature*, **356**, 408–414.

106. Sharp,L.L., Zhou,J. and Blair,D.F. (1995) Tryptophan-scanning mutagenesis of MotB, an integral membrane protein essential for flagellar rotation in Escherichia coli. *Biochemistry*, **34**, 9166–9171.

107. Rasmussen,T., Rasmussen,A., Singh,S., Galbiati,H., Edwards,M.D., Miller,S. and Booth,I.R. (2015) Properties of the Mechanosensitive Channel MscS Pore Revealed by Tryptophan Scanning Mutagenesis. *Biochemistry*, **54**, 4519–4530.

108. Depriest,A., Phelan,P. and Martha Skerrett,I. (2011) Tryptophan scanning mutagenesis of the first transmembrane domain of the innexin Shaking-B(Lethal). *Biophys. J.*, **101**, 2408–2416.

109. Weinglass,A.B., Smirnova,I.N. and Kaback,H.R. (2001) Engineering conformational flexibility in the lactose permease of Escherichia coli: use of glycine-scanning mutagenesis to rescue mutant Glu325-->Asp. *Biochemistry*, **40**, 769–776.

110. Olson,C.A., Wu,N.C. and Sun,R. (2014) A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. *Curr. Biol.*, **24**, 2643–2651.

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

112. Vigneri,R., Squatrito,S. and Sciacca,L. (2010) Insulin and its analogs: actions via insulin and IGF receptors. *Acta Diabetol.*, **47**, 271–278.

113. Karczewski,K.J., Weisburd,B., Thomas,B., Solomonson,M., Ruderfer,D.M., Kavanagh,D., Hamamsy,T., Lek,M., Samocha,K.E., Cummings,B.B., *et al.* (2017) The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucl. Acids Res.*, **45**, D840–D845.

114. Zou,J., Valiant,G., Valiant,P., Karczewski,K., Chan,S.O., Samocha,K., Lek,M., Sunyaev,S., Daly,M. and MacArthur,D.G. (2016) Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nat. Commun.*, **7**, 13293.

115. MacArthur,D.G., Manolio,T.A., Dimmock,D.P., Rehm,H.L., Shendure,J., Abecasis,G.R., Adams,D.R., Altman,R.B., Antonarakis,S.E., Ashley,E.A., *et al.* (2014) Guidelines for investigating causality of sequence variants in human disease. *Nature*, **508**, 469–476.

116. Gasperini,M., Starita,L. and Shendure,J. (2016) The power of multiplexed functional analysis of genetic variants. *Nat. Protoc.*, **11**, 1782–1787.

117. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

118. Stenson,P.D., Ball,E.V., Mort,M., Phillips,A.D., Shaw,K. and Cooper,D.N. (2012) The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Hum. Genet.*, **133**, 1–9.

119. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human

genes and genetic disorders. *Nucl. Acids Res.*, **43**, 789–798.

120. Landrum,M.J., Lee,J.M., Riley,G.R., Jang,W., Rubinstein,W.S., Church,D.M. and Maglott,D.R. (2013) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucl. Acids Res.*, **44**, 862–868.

121. Wan,P.T.C., Garnett,M.J., Roe,S.M., Lee,S., Niculescu-Duvaz,D., Good,V.M., Jones,C.M., Marshall,C.J., Springer,C.J., Barford,D., *et al.* (2004) Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell*, **116**, 855–867.

122. Rodriguez-Viciana,P., Tetsu,O., Tidyman,W.E., Estep,A.L., Conger,B.A., Cruz,M.S., McCormick,F. and Rauen,K.A. (2006) Germline mutations in genes within the MAPK pathway cause cardio-facio-cutaneous syndrome. *Science*, **311**, 1287–1290.

123. Mester,J. and Eng,C. (2013) When overgrowth bumps into cancer: the PTEN-opathies. *Am. J. Med. Genet.*, **163**, 114–121.

124. Matteucci,M.D. and Heyneker,H.L. (1983) Targeted random mutagenesis: the use of ambiguously synthesized oligonucleotides to mutagenize sequences immediately 5' of an ATG initiation codon. *Nucl. Acids Res.*, **11**, 3113–3121.

125. Kato,S., Han,S.-Y., Liu,W., Otsuka,K., Shibata,H., Kanamaru,R. and Ishioka,C. (2003) Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proc. Natl. Acad. Sci.*, **100**, 8424–8429.

126. Kumar,S., Suleski,M.P., Markov,G.J., Lawrence,S., Marco,A. and Filipski,A.J. (2009) Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations. *Genome Res.*, **19**, 1562–1569.

127. Saunders,C.T. and Baker,D. (2002) Evaluation of Structural and Evolutionary Contributions to Deleterious Mutation Prediction. *J. Mol. Biol.*, **322**, 891–901.

128. Welsh,M.J. and Smith,A.E. (1993) Molecular Mechanisms of CFTR Chloride Channel Dysfunction in Cystic Fibrosis. *Cell*, **73**, 1251–1254.

129. Seakins,M., Gibbs,W.N., Milner,P.F. and Bertles,J.F. (1973) Erythrocyte Hb-S concentration. An important factor in the low oxygen affinity of blood in sickle cell anemia. *J. Clin. Invest.*, **52**, 422–432.

130. Sunyaev,S.R., Eisenhaber,F., Rodchenkov,I.V., Eisenhaber,B., Tumanyan,V.G. and Kuznetsov,E.N. (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.*, **12**, 387–394.

131. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlić,A., Quesada,M., Quinn,G.B., Westbrook,J.D., *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucl. Acids Res.*, **39**, 392–401.

132. Chiti,F. and Dobson,C.M. (2017) Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade. *Annu. Rev. Biochem.*, **86**, 27–68.

133. Balbach,J.J., Petkova,A.T., Oyler,N.A., Antzutkin,O.N., Gordon,D.J., Meredith,S.C. and Tycko,R. (2002) Supramolecular structure in full-length Alzheimer's beta-amyloid fibrils: evidence for a parallel beta-sheet organization from solid-state nuclear magnetic resonance. *Biophys. J.*, **83**, 1205–1216.

134. Wagenaar,T.R., Ma,L., Roscoe,B., Park,S.M., Bolon,D.N. and Green,M.R. (2014) Resistance to vemurafenib resulting from a novel mutation in the BRAFV600E kinase domain. *Pigment Cell & Melanoma Research*, **27**, 124–133.

135. Landrum,M.J., Lee,J.M., Benson,M., Brown,G., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Hoover,J., *et al.* (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucl. Acids Res.*, **44**, 862–868.

136. García-Nafría,J., Watson,J.F. and Greger,I.H. (2016) IVA cloning: A single-tube universal cloning system exploiting bacterial In Vivo Assembly. *Sci Rep*, **6**, 27459.

137. Grabowski,T.J., Cho,H.S., Vonsattel,J.P., Rebeck,G.W. and Greenberg,S.M. (2001) Novel amyloid precursor protein mutation in an Iowa family with dementia and severe cerebral amyloid angiopathy. *Ann. Neurol.*, **49**, 697–705.

138. Shimizu,T., Fukuda,H., Murayama,S., Izumiyama,N. and Shirasawa,T. (2002) Isoaspartate formation at position 23 of amyloid beta peptide enhanced fibril formation and deposited onto senile plaques and vascular amyloids in Alzheimer's disease. *J. Neurosci. Res.*, **70**, 451–461.

139. Van Nostrand,W.E., Melchor,J.P., Cho,H.S., Greenberg,S.M. and Rebeck,G.W. (2001) Pathogenic effects of D23N Iowa mutant amyloid beta -protein. *J. Biol. Chem.*, **276**, 32860–32866.

140. Crnkovic,I. (2002) Component-based Software Engineering. *Software Focus*, **2**, 1–7.

141. Baker,D. and Sali,A. (2001) Protein Structure Prediction and Structural Genomics. *Science*, **294**, 93–96.

142. Stein,R.R., Marks,D.S. and Sander,C. (2015) Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. *PLoS Comput. Biol.*, **11**, e1004182.

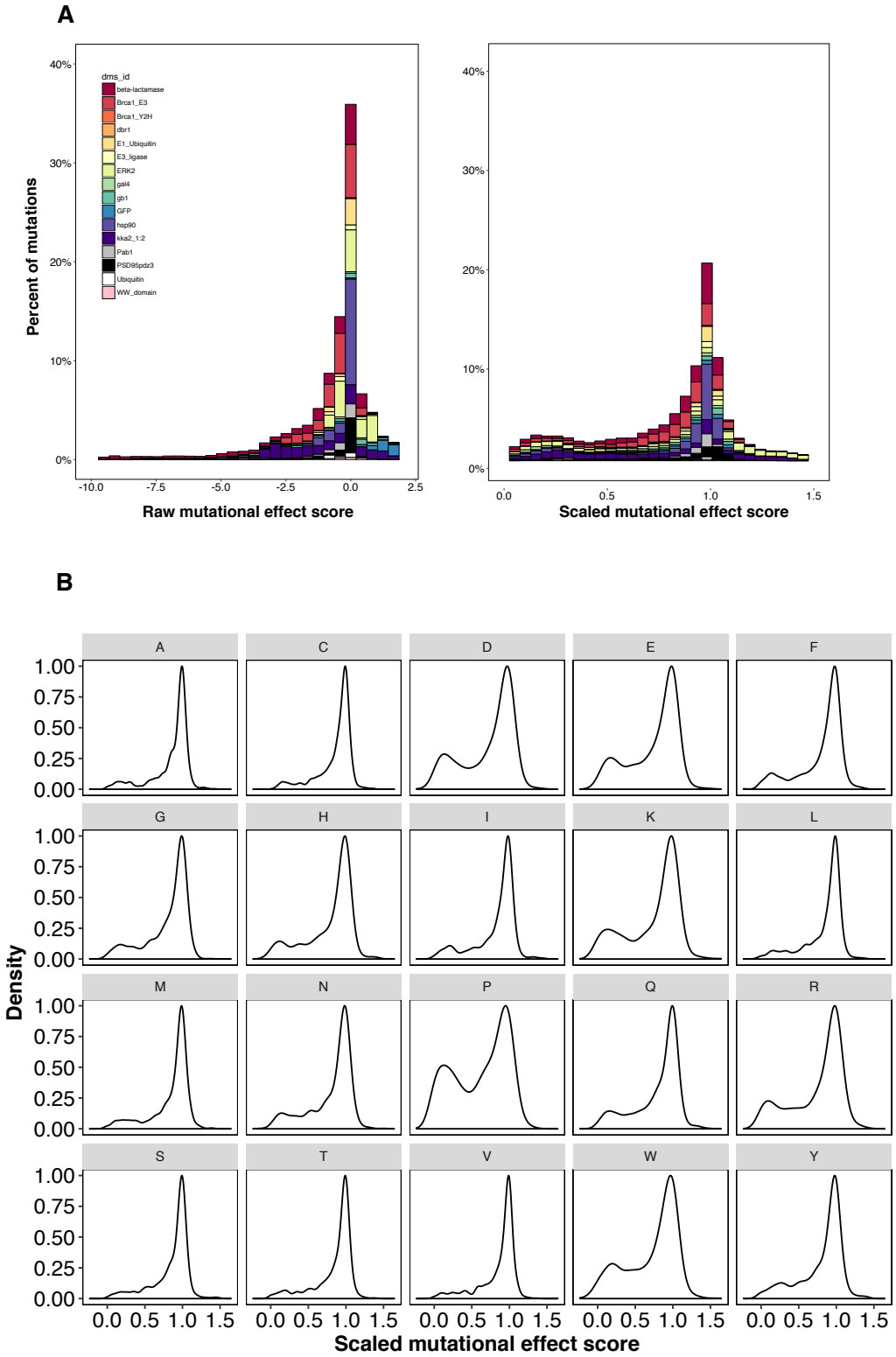
143. Sheridan,R., Fieldhouse,R.J., Hayat,S., Sun,Y., Antipin,Y., Yang,L., Hopf,T., Marks,D.S. and Sander,C. (2015) EVfold.org: Evolutionary Couplings and Protein 3D Structure Prediction. *bioRxiv*, 10.1101/021022.

144. Lakhani,B., Thayer,K.M., Hingorani,M.M. and Beveridge,D.L. (2017) Evolutionary Covariance Combined with Molecular Dynamics Predicts a Framework for Allostery in the MutS DNA Mismatch Repair Protein. *J Phys Chem B*, **121**, 2049–2061.

145. Kamisetty,H., Ovchinnikov,S. and Baker,D. (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci.*, **110**, 15674–15679.

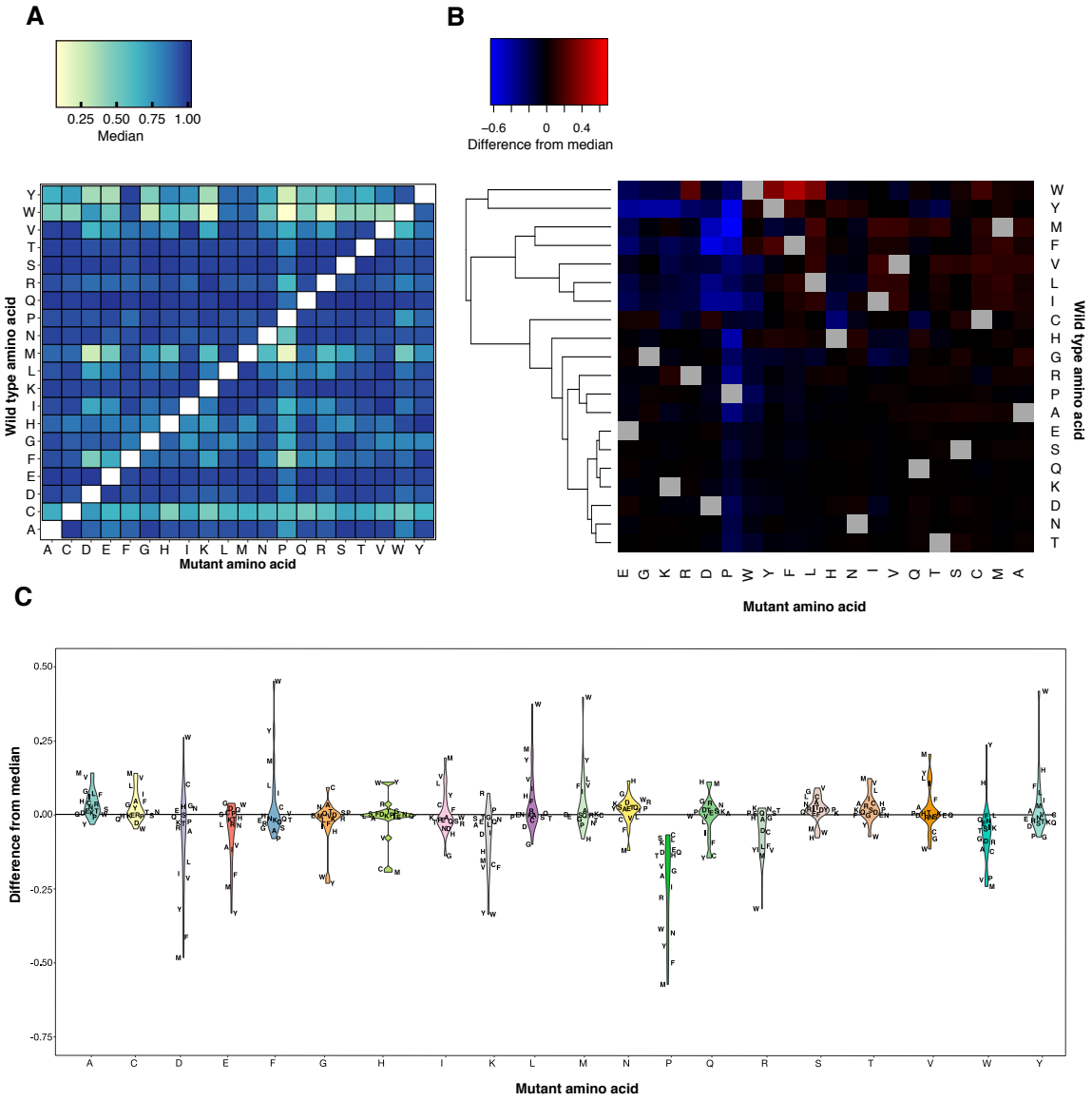
146. Salinas,V.H. and Ranganathan,R. (2017) Inferring amino acid interactions underlying protein function. *bioRxiv*, 10.1101/215368.

APPENDIX A

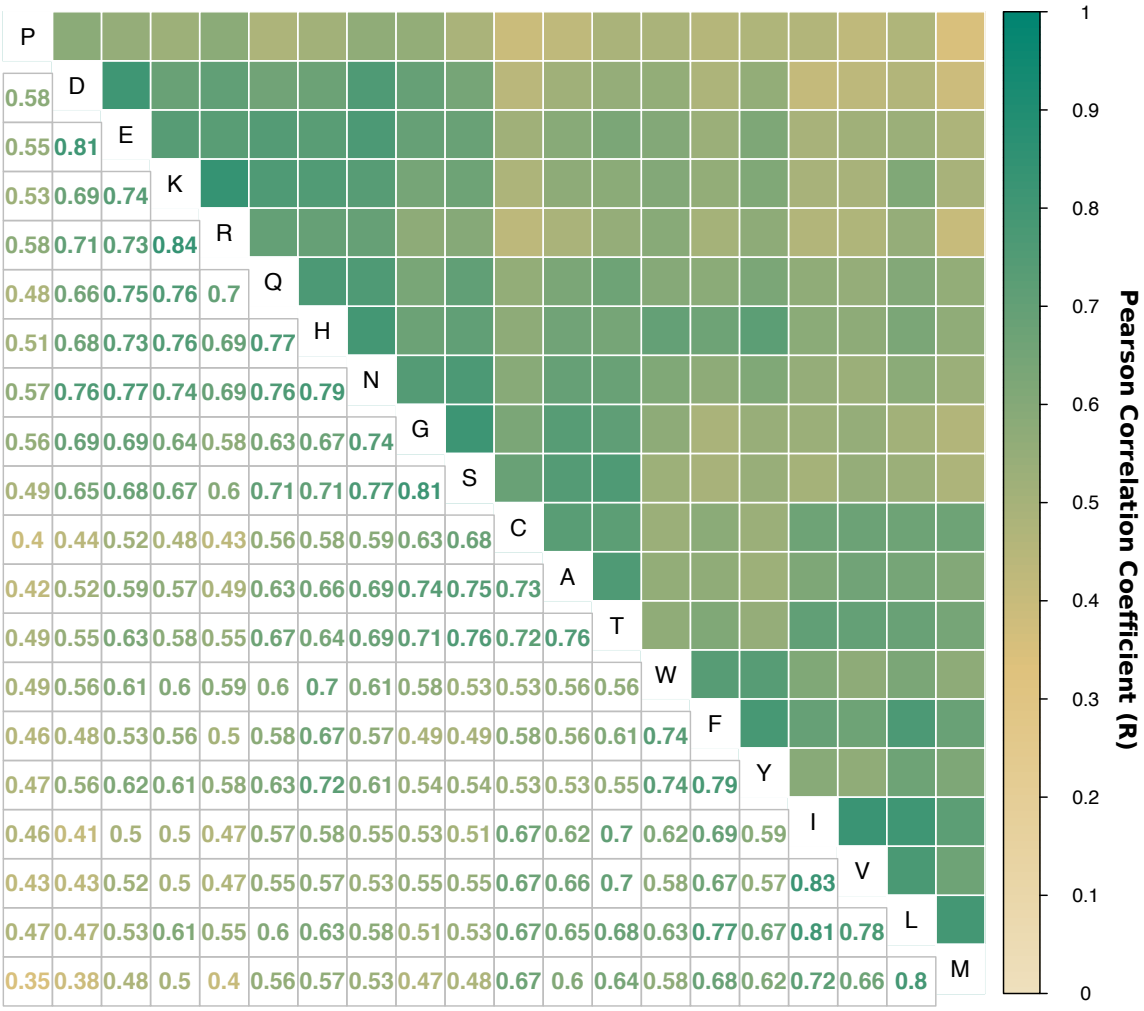


Appendix 1 I curated large-scale mutagenesis data sets describing the effects of 34,373 mutations at 2,236 positions in fourteen proteins. To facilitate comparisons between each data

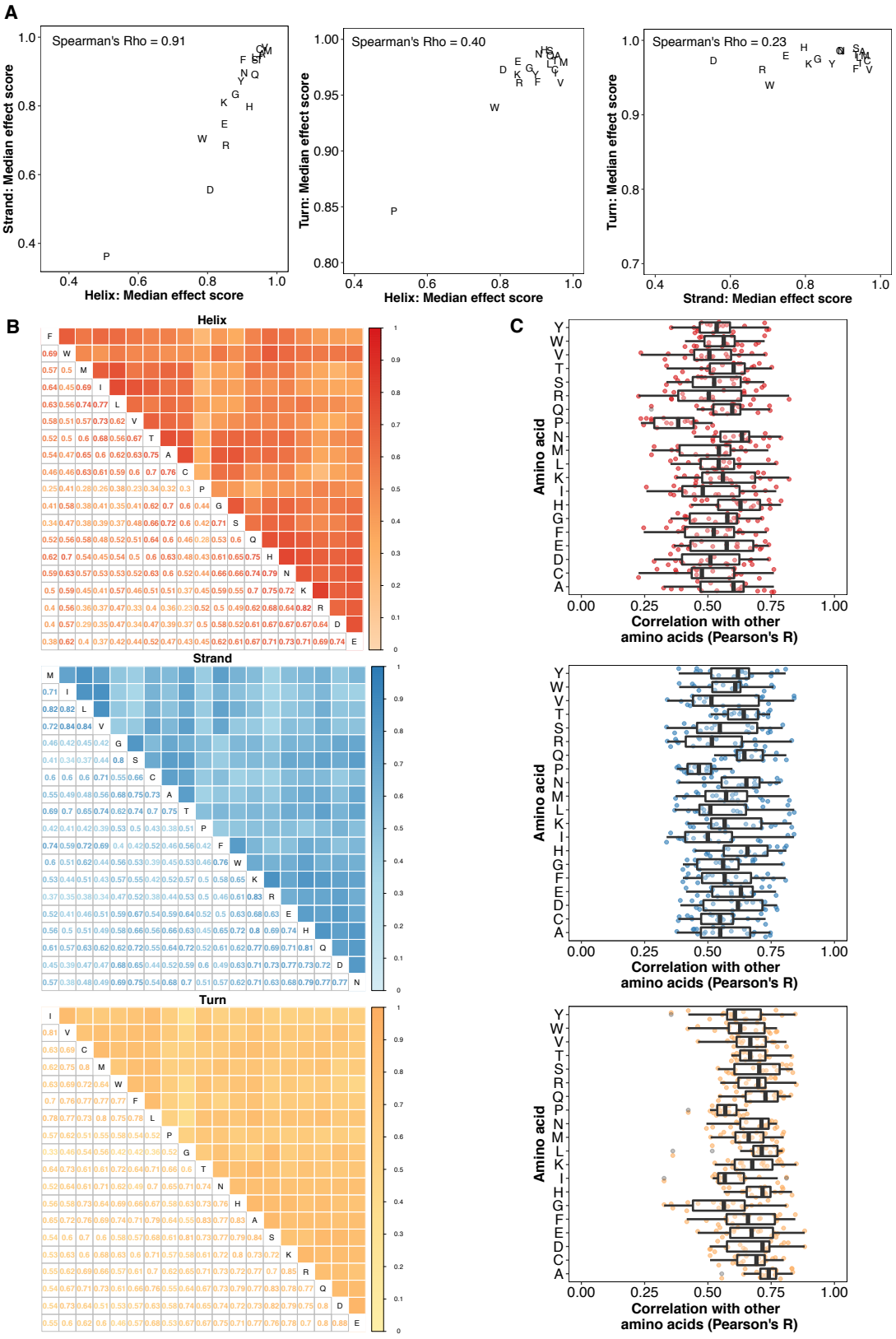
set, I rescaled mutational effect scores for each protein by subtracting the median mutational effect score of all synonymous mutations in that protein from each nonsynonymous mutational effect score and then dividing that difference by the median of the bottom 1% of mutational effect scores. (**A**) Stacked histograms of the original scores (**left panel**) and rescaled scores (**right panel**) are shown. (**B**) Density plots of the scaled mutational effect scores for each amino acid substitution are shown.



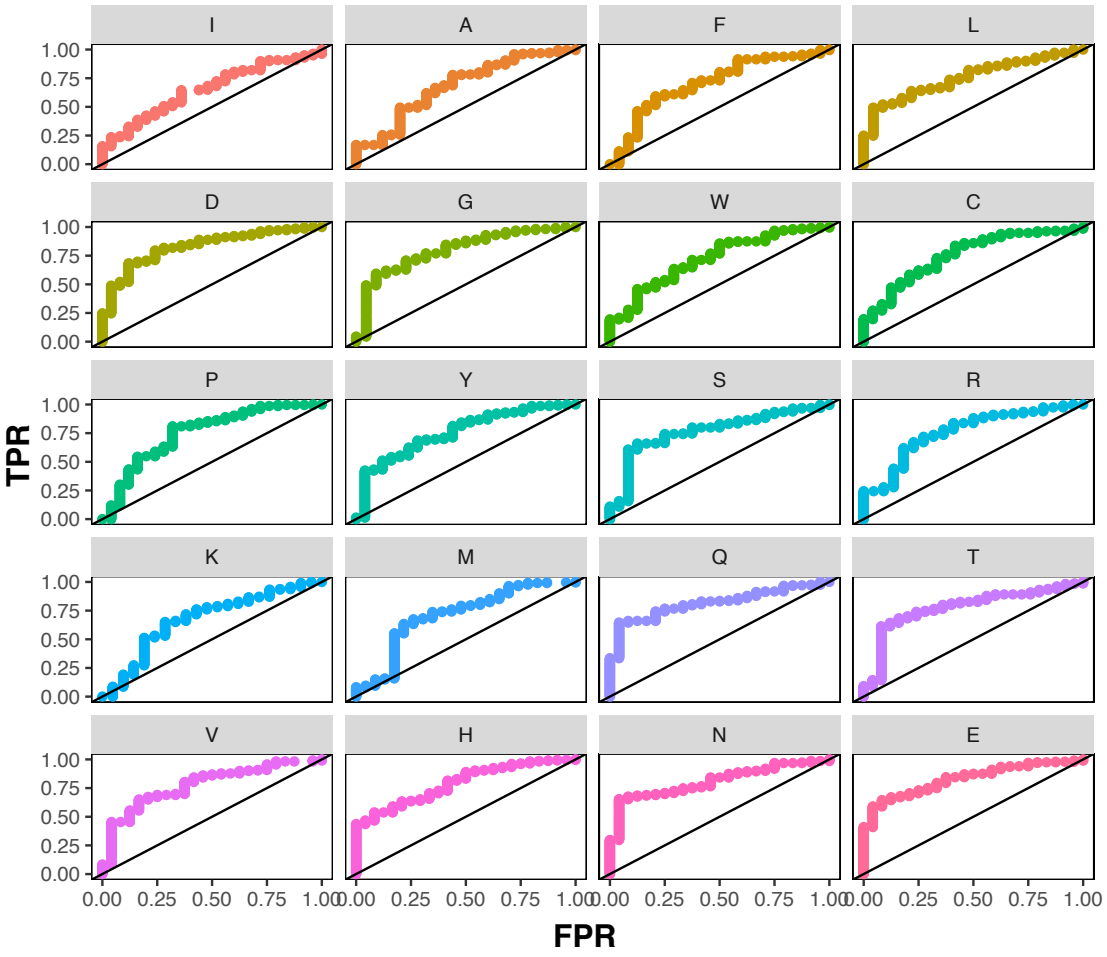
Appendix 2 (**A**) The median effect of each substitution for every wild type amino acid is presented in a heat map where substitutions with disruptive effects are shown in yellow and wild type-like effects are shown in blue. (**B**) For each substitution and wild type amino acid pair, I subtracted the median effect of all substitutions at positions with the wild type amino acid from the median effect of the substitution at those positions. The differences between these medians are shown in a heat map. A difference greater than zero, shown in red, denotes a substitution that was more tolerated than the median substitution for that wild type amino acid, while a difference less than zero, shown in blue, denotes a more disruptive substitution. I used hierarchical clustering (linkage method: complete) on the median differences to cluster both the wild type and mutant amino acids. **C**) The differences between medians, calculated as described for panel B, are shown in violin plots. Each violin plot reveals the dependence of the effect of a substitution on the identity of the wild type amino acid (labeled on the plot).



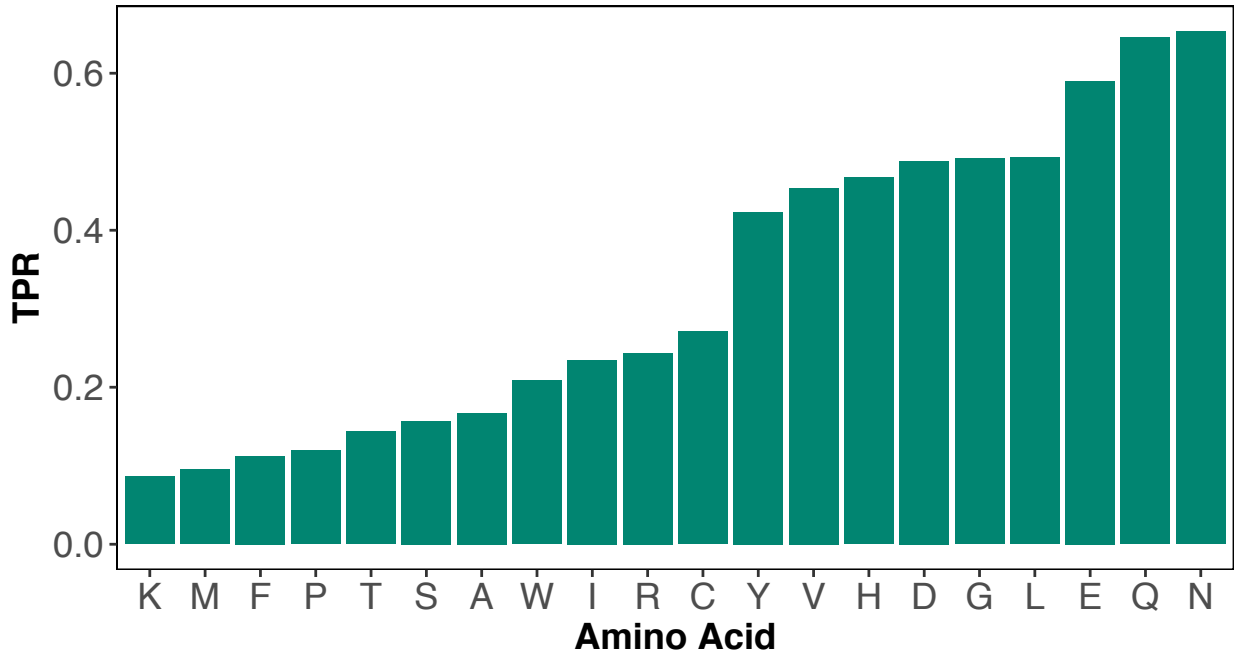
Appendix 3 For each substitution, Pearson correlation coefficients were calculated for the mutational effect scores of that substitution with every other substitution at each position. A correlation plot of these Pearson coefficients is shown. Color indicates the Pearson correlation coefficient ranging from 0 (light brown) to 1 (green).


A
Spearman's Rho = 0.91
Strand: Median effect score
Helix: Median effect score
Spearman's Rho = 0.40
Turn: Median effect score
Helix: Median effect score
Spearman's Rho = 0.23
Turn: Median effect score
Strand: Median effect score
B
Helix
Strand
Turn
C
Amino acid
Correlation with other amino acids (Pearson's R)

Appendix 4 (**A**) For each amino acid substitution, the median mutational effect score was calculated. The correlation between the median mutational effects for each substitution in helices, strand and turns are shown in scatterplots, and Spearman's Rho indicates the degree of rank correlation within each scatterplot. (**B**) Pearson correlation coefficients were calculated for the mutational effects of each substitution with every other substitution at every position. The Pearson correlation coefficient plots are shown separately for α-helices (top), β-sheets (middle), and turns (bottom). (**C**) Boxplots show the distribution of Pearson correlation coefficients for each amino acid type in three structural contexts.

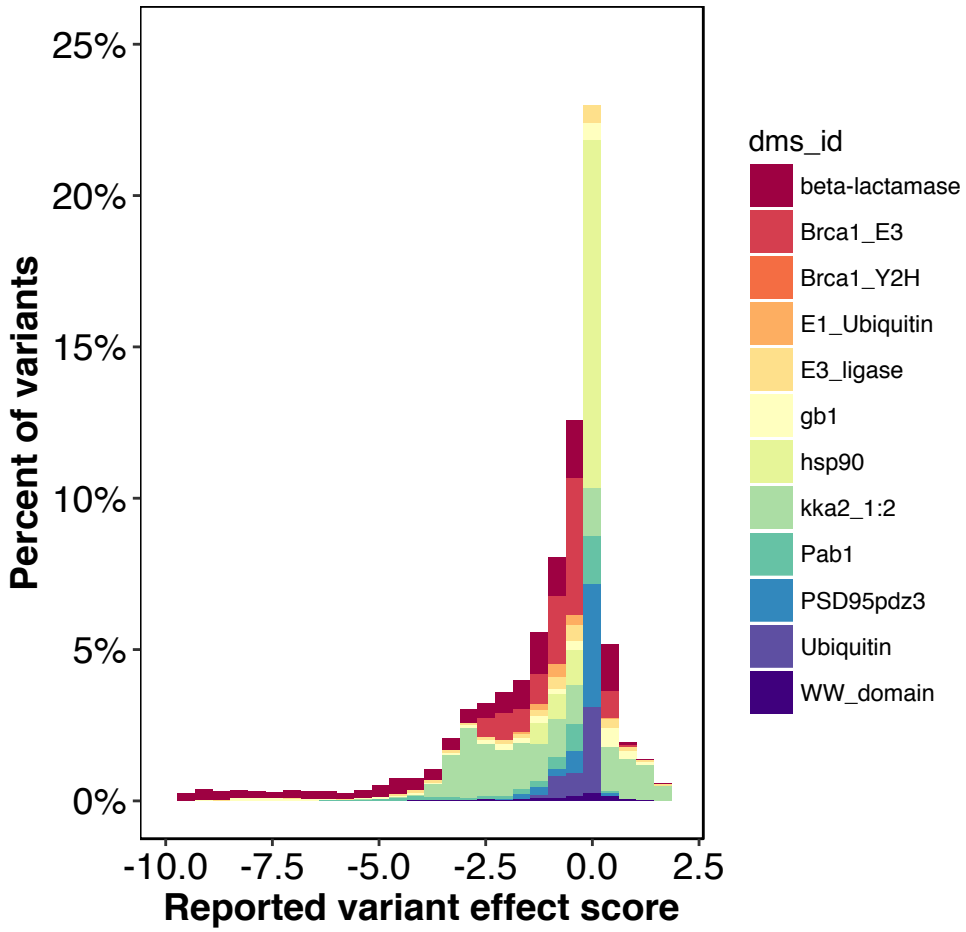


Appendix 5 A mutational effect threshold was defined such that positions with a mutational effect score below the threshold were classified as "interface," whereas positions with a mutational effect score above the threshold were classified as "non-interface." ROC curves were generated by varying this threshold for each amino acid type in the four proteins with protein or DNA ligand-bound structures (hYAP65 WW domain, PSD95 pdz3 domain, Gal4 and BRCA1 RING domain (BARD1 binding)).

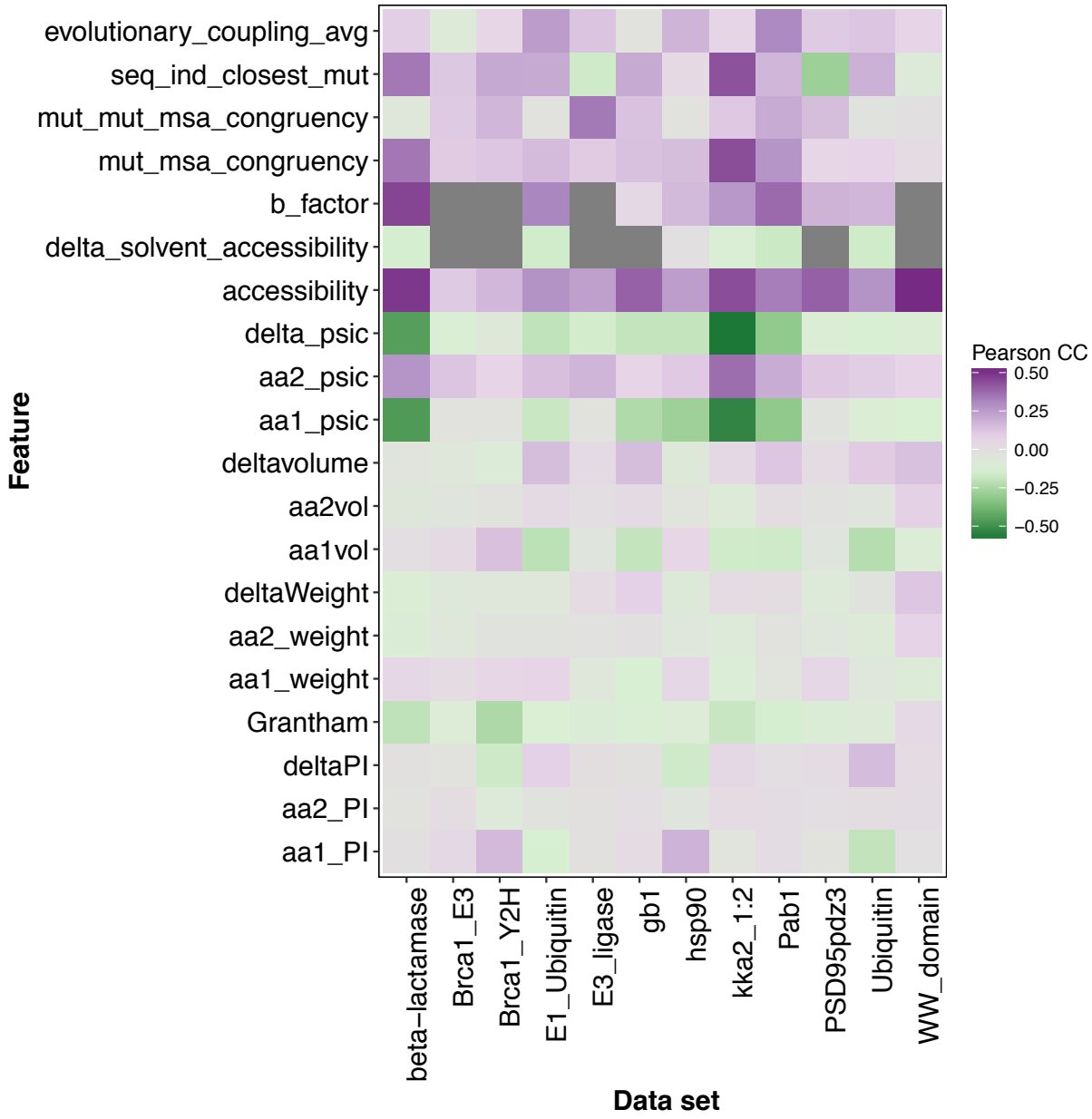


Appendix 6 A mutational effect threshold was defined such that positions with a mutational effect score below the threshold were classified as "interface," whereas positions with a mutational effect score above the threshold were classified as "non-interface." A barplot shows each amino acid substitution's true positive rate (TPR) for detecting interface positions at a fixed, 5% non- interface position false positive rate.

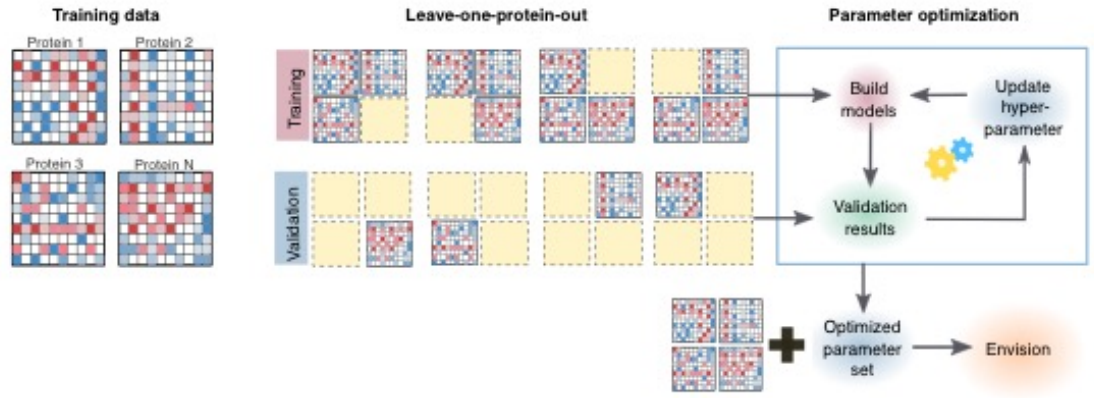
APPENDIX B



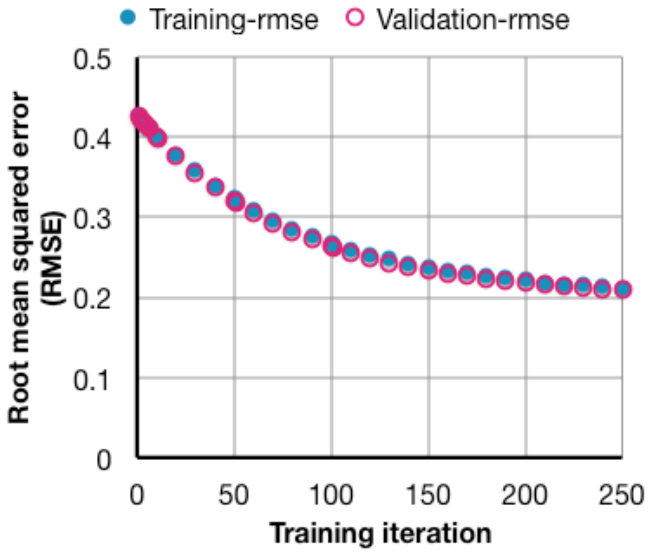
Appendix B 1 A histogram shows the distributions of reported variant effect scores from 12 large-scale mutagenesis data sets.



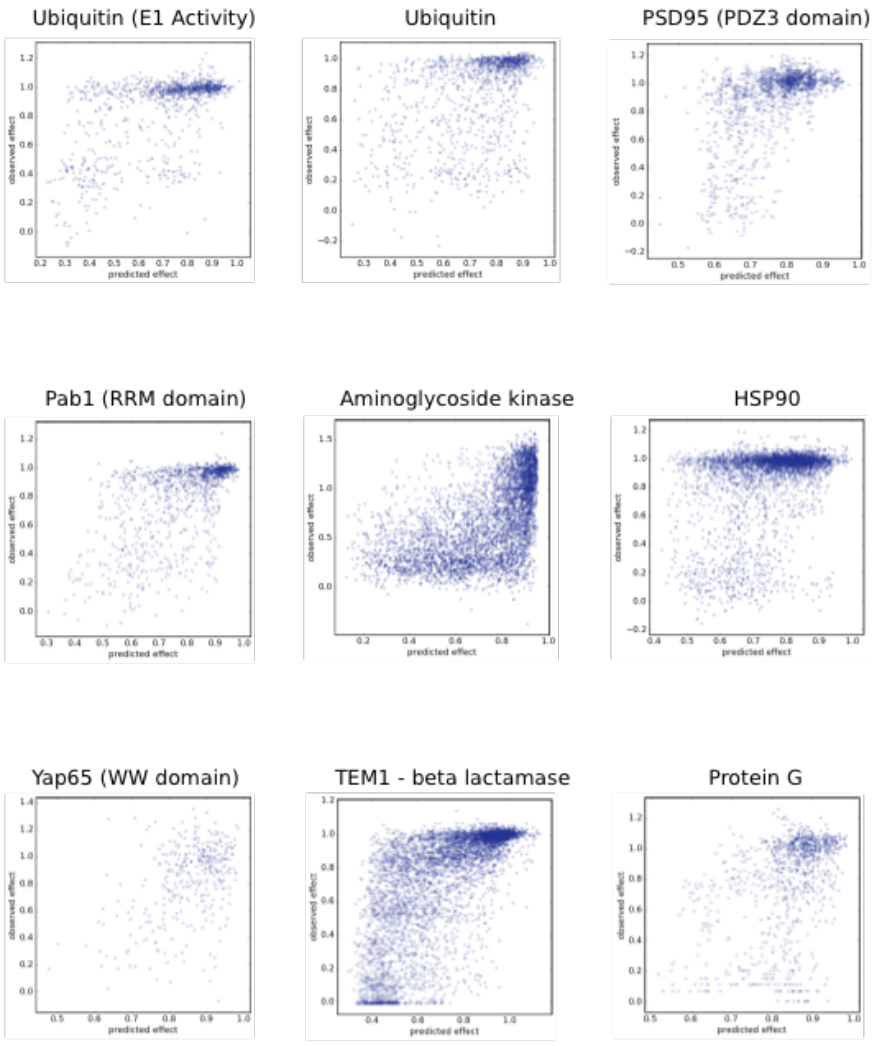
Appendix B 2 A heatmap shows the Pearson correlation coefficient between descriptive feature values and variant effect scores for each large-scale mutagenesis data set. Note, E3 ligase, and BRCA1 datasets are missing B factor and predicted change in solvent accessibility features and also have low correlations between existing features and effect scores.



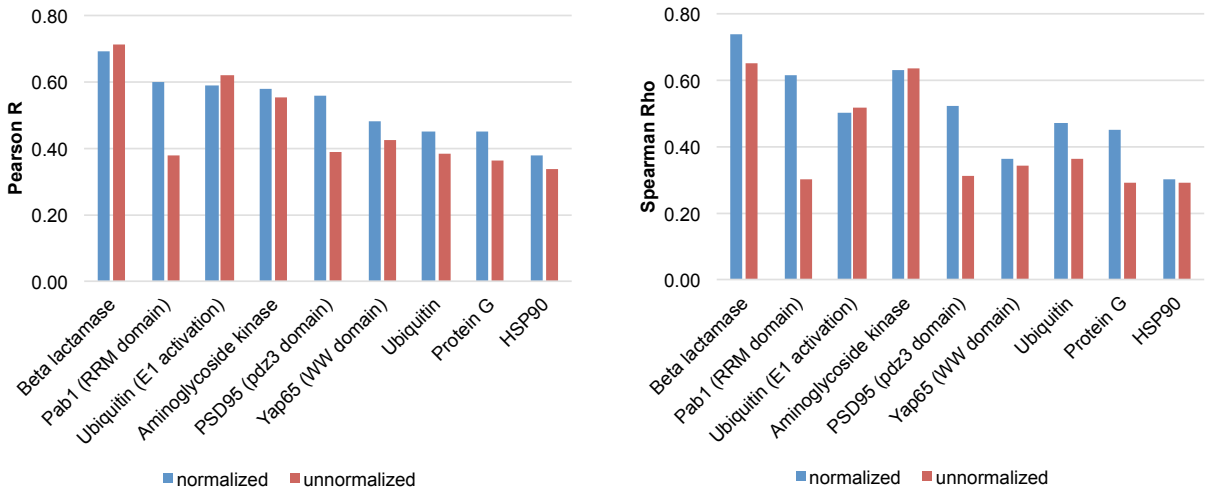
Appendix B 3 Our hyperparameter tuning scheme is designed to generate generalizable models. To determine the optimal values for each hyperparameter, I used a leave-one-protein-out cross-validation approach. To begin, I collected large-scale mutagenesis data sets and annotated them with features. Next, I created 8 training and validation dataset pairs; each training set contains variants from 7 of 8 proteins and the validation set contains variants from the protein withheld from the training set. Thus, each parameter set is being evaluated for its ability to predict a protein unseen by the model. Then, I test a set of hyperparameters using all testing and validation pair sets, and then update hyperparameters until all parameter values are evaluated. Once completed, I identify the parameter set that yields the most generalizable model, i.e., performs best on the left out protein's variant data set.



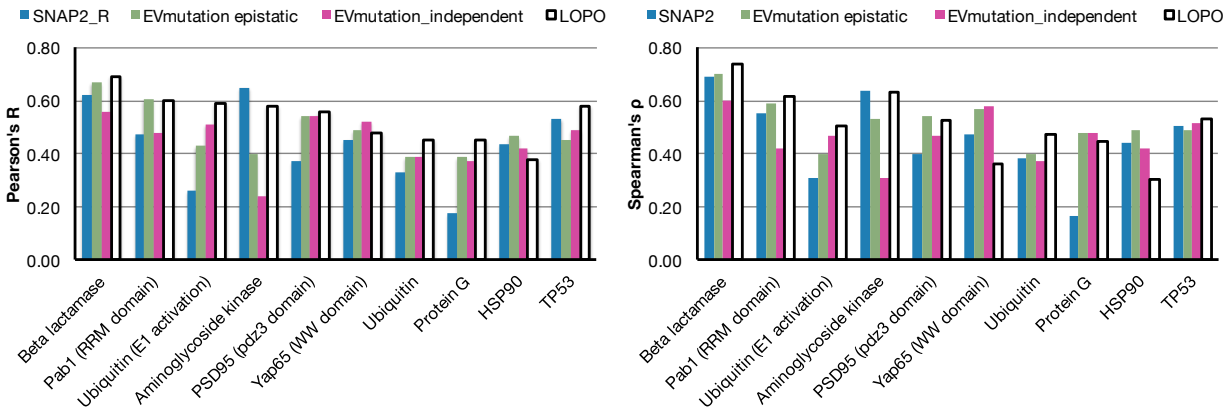
Appendix B 4 Training and testing data set RMSEs are very similar across iterations. While training Envision, 5% of data was withheld to determine the performance of the model as each tree was trained and added to the ensemble of decision trees. The plot shows the root mean squared error (RMSE), otherwise known as the mean difference between observed and predicted scores, for training and validation data. There is little difference between the RMSE of Envision for training and testing data, which suggests that Envision is not over trained.



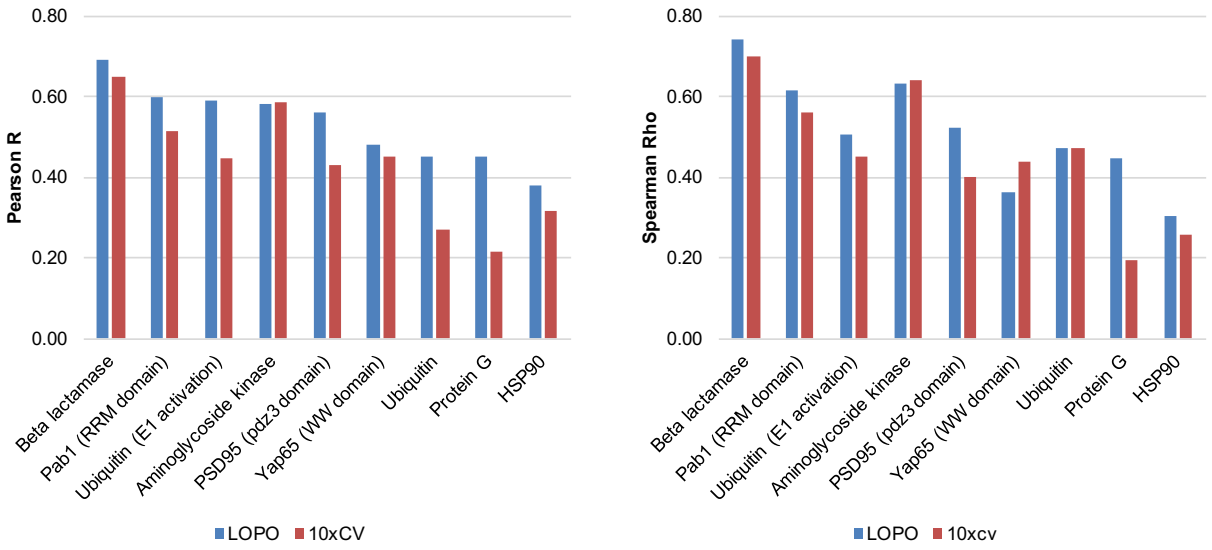
Appendix B 5 Scatter plots show the correlation between leave-one-protein-out model predictions and observed variant effects.



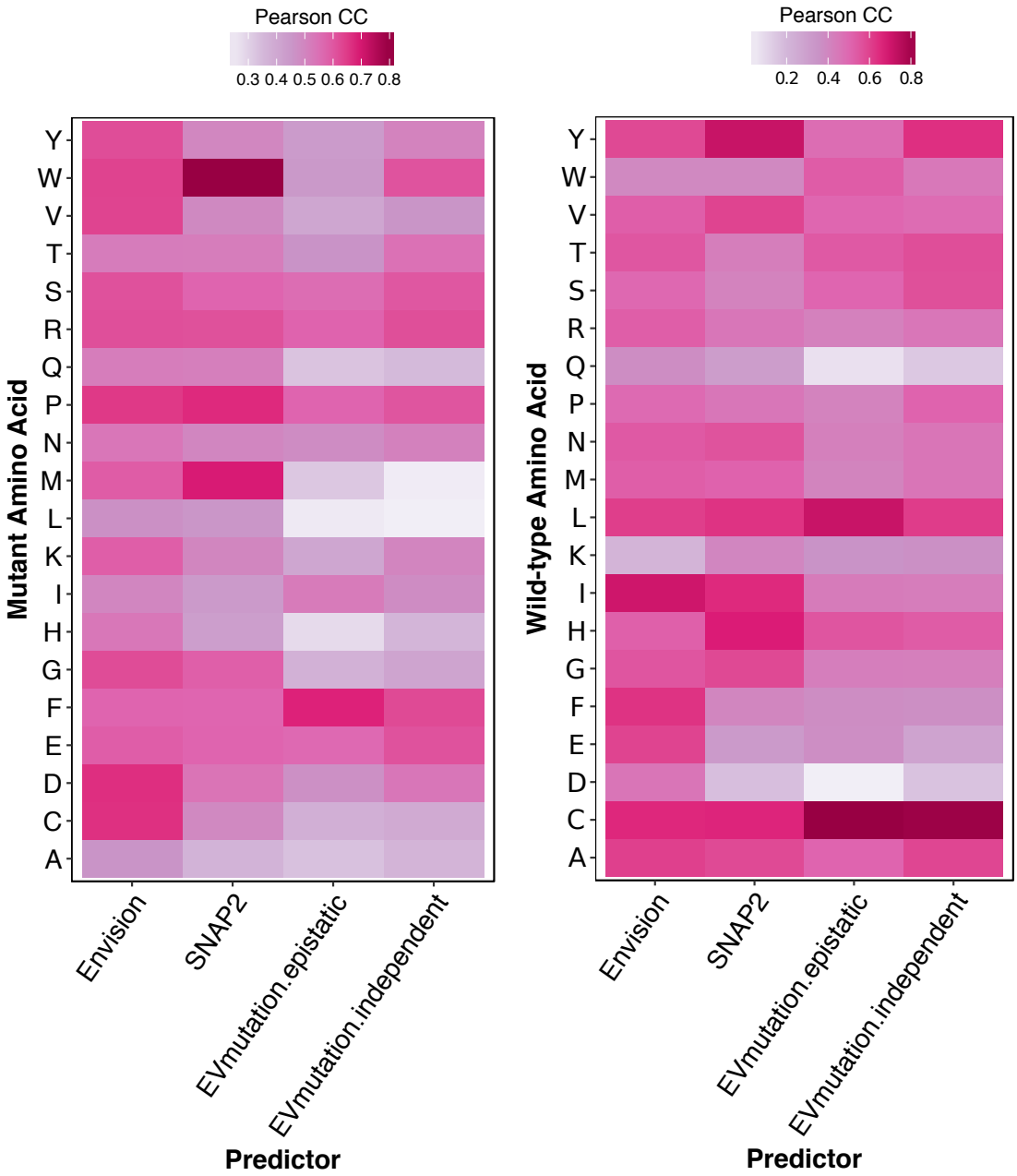
Appendix B 6 Leave-one-protein-out models were trained either with normalized or non-normalized variant effect scores. The barplots show Pearson's (left) and Spearman's (right) correlation coefficients between observed variant effect scores and predicted variant effect scores for the left-out protein from models trained using normalized (blue) or non-normalized (red) scores. Overall, models trained on normalized variant effect scores predicted the left-out protein variant effect scores best.



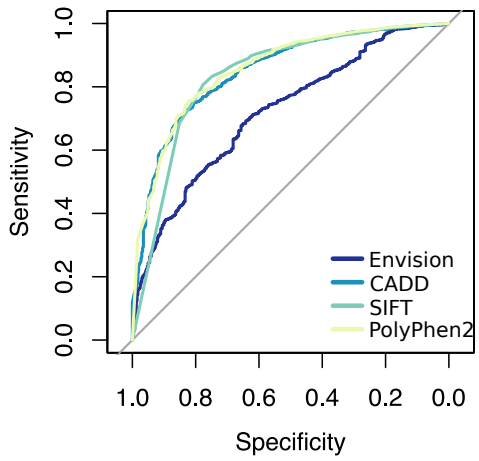
Appendix B 7 Our leave-one-protein-out models compare favorably to SNAP2 and EVmutation models. This barplot shows the correlation between predicted and observed variant effect scores for each data set for SNAP2, EVmutation (epistatic and independent models) and our leave-one-protein-out models. The x-axis shows the protein/domain withheld from training. Here, I observe that our models outperform other predictors that our models have yet to see in training.



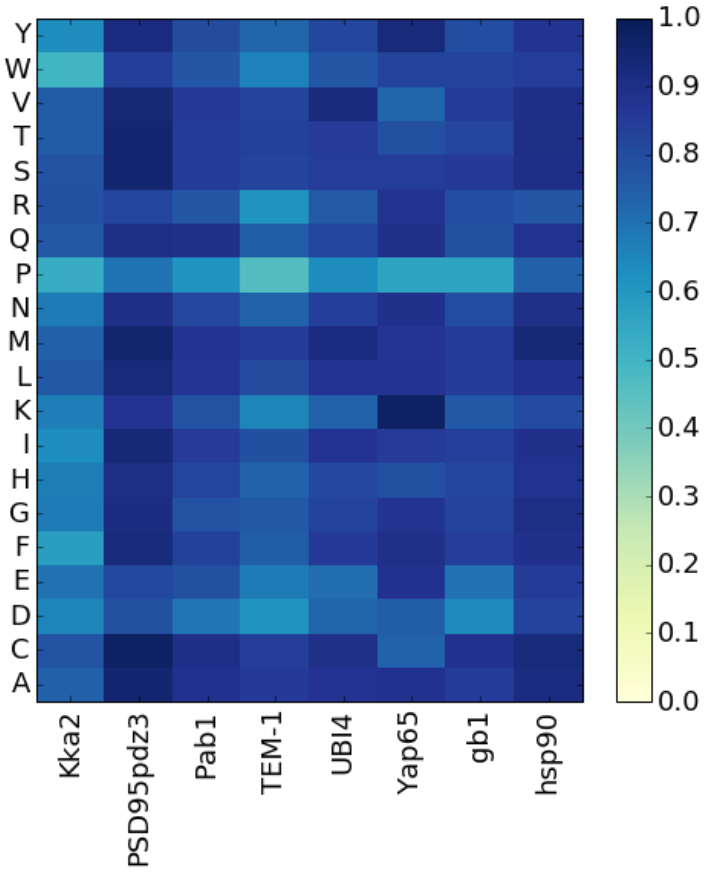
Appendix B 8 Effect of hyperparameter tuning cross-validation procedure. These barplots show the Pearson (left) and Spearman (right) correlations (y-axes) between predicted and observed variant effect scores for the left-out protein for models trained with hyperparameters optimized using a leave-one-protein-out cross-validation approach (blue). In this approach, at each round of cross-validation a different protein was used for testing. A standard tenfold cross-validation was also tested, where at each round of cross-validation a random 10% of variant effect scores were used for testing (red). The x-axes show the protein or domain left out of the hyperparameter tuning and model training procedures, which was used to evaluate model performance.



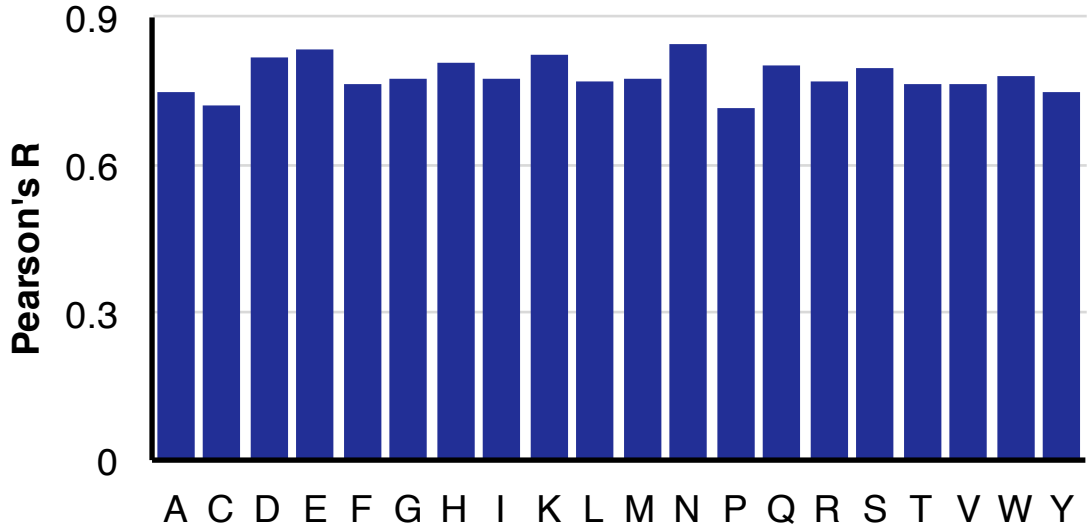
Appendix B 9 Heatmaps show the correlation (Pearson's R) between predictions from four predictors for TP53 mutations arcoss mutant (**A**) and wild-type (**B**) amino acids. Darker red denotes more accurate predictions, while white shows poor predictive performance.



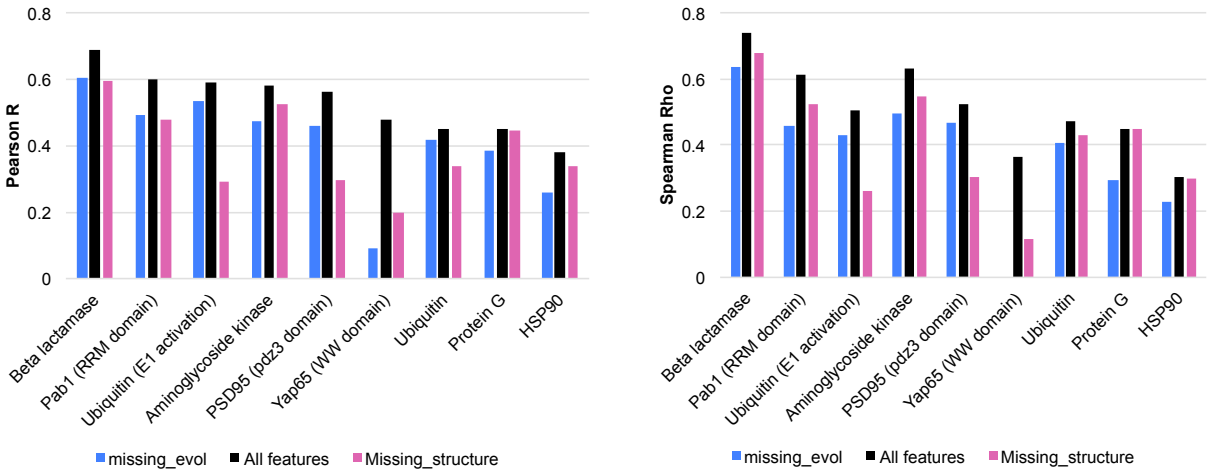
Appendix B 10. Envision, CADD, SIFT and PolyPhen2 were used to predict 9,028 pathogenic and 402 benign mutations from the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/). Receiver operator characteristic (ROC) curves were generated for each model using the pROC package in R. PolyPhen2 predicted pathogenicity with the highest accuracy (AUC = 0.86, 95% CI: 0.84-0.88) followed by CADD (0.85, 0.83-0.87), SIFT (0.84, 0.81-0.86) and then Envision (0.72, 0.70-0.74). Confidence intervals were determined with 2,000 bootstrap replicates.



Appendix B 11 The heatmap below shows the mean variant effect score for each of the twenty amino acids across eight protein data sets. It is clear that proline mutations are one of the most disruptive mutations to protein function.



Appendix B 12 A barplot shows the correlation between Envision predictions and observed variant effect scores for each mutant amino acid in our training data. The mutant amino acid type is shown on the x-axis.



Appendix B 13 The leave-one-protein-out models I trained were used to predict their left-out protein's variant effect scores with one of three different feature sets. The barplots above show Pearson's (left) and Spearman's (right) correlation coefficients between predicted variant effect scores and observed variant effect scores for each of the left-out proteins. Black bars indicate that all features were used during the prediction phase (i.e. the same data as Figure 3B). Pink bars denote predictions made when all structural features for the left-out protein were masked. Blue bars denote predictions made when all evolutionary conservation-related features were masked. Structural features are identified in green in Figure 2D, and evolutionary features are identified in blue in Figure 2D.

Appendix B Table 1 Summary of large-scale mutagenesis datasets.

Name	protein	dms_id	first_author	PMID	Year	Region mutagenized	Number of mutants	Number of mutagenized protein positions	Organism	Selected phenotype	UniProt_ID	PDB_ID	Replicate correlation	Used in model?	Molecular function	Structural folds
TEM1 β-lactamase	TEM1 β-lactamase	Beta lactamase	Firnberg	24567513	2014	Full protein	5198	287	*E. coli*	Ampicillin resistance	P62593	1XPB	?	YES	hydrolysis of lactam antibiotics	Helix, sheet, turn
Yap65 (WW domain)	Yap65	WW_domain	Fowler	20711194	2010	WW domain	363	34	*H. sapiens*	Substrate binding	P46937	1JMQ	NA	YES	Protein binding	Beta, turn
PSD95 (Pdz3 domain)	PSD95	PSD95pdz3	McLaughlin	23041932	2012	PDZ3 domain	1577	83	*Rattus norvegicus*	Ligand binding	P31016	2BE9	NA	YES	Protein kinase bining	helix, sheet
Brca1 (RING domain)-E3 ligase activity	Brca1	Brca1_E3	Starita	25823446	2015	RING1 domain	4872	303	*H. sapiens*	Ubiquitin ligase activity	P38398	1JM7	~0.85	NO	Many	Helix, sheet, turn
Brca1 (RING domain)-Bard1 binding	Brca1	Brca1_Y2H	Starita	25823446	2015	RING1 domain	1748	102	*H. sapiens*	Binding activity (Y2H)	P38398	1JM7	~0.85	NO	Many	Helix, sheet, turn
Aminoglycoside kinase	Aminoglycoside kinase	kka2_1:2	Melnikov	24914046	2014	Full protein	5300	264	*K. pneumoniae*	Antibiotic resistance	P00552	1ND4	0.88	YES	Kanamycin kinase activity	Helix, sheet, turn
E4B (U-box domain)	E4B (U-box domain)	E3_ligase	Starita	23509263	2013	U-box domain	899	102	*M. musculus*	Ubiquitin ligase activity	Q9ESQ0	2KR4	0.94	NO	Ubiquitin activating enzyme activity	Helix, sheet, turn
Hsp90	Hsp90	hsp90	Mishra	27068472	2016	N/A	4021	219	*S. cerevisiae*	Yeast growth	P02829	2CG9	0.96	YES	Unfolded protein binding	Helix, sheet, turn
Ubiquitin	Ubiquitin	Ubiquitin	Roscoe	23376099	2013	Full peptide	1249	75	*S. cerevisiae*	Yeast growth rate	P0CG63	3CMM	0.96	YES	ATP-dependent protein binding	Helix, sheet, turn
Pab1 (RRM domain)	Pab1	Pab1	Melamed	25671604	2013	RRM domain	1188	75	*S. cerevisiae*	mRNA binding	P04147	1CVJ	NA	YES	Poly-A binding	Helix, sheet, turn
Ubiquitin - E1 activity	Ubiquitin	E1_Ubiquitin	Roscoe	24862281	2014	N/A	1085	60	*S. cerevisiae*	Yeast growth	P0CG63	3CMM	0.98	YES	ATP-dependent protein binding	Helix, sheet, turn
Protein G (IgG domain)	Protein G	gb1	Olson	25455030	2014	IgG-binding domain	1045	55	*Streptococcus sp. group G*	IgG-Fc binding	P06654	1PGA	0.99	YES	IgG-binding	helix, sheet

Appendix B Table 2 Summary of descriptive features used to train gradient boosted models.

Features	Name	Description	Range/Categories	Reference
AA1	WT amino acid	WT AA	All possible AA	NA
AA2	MT amino acid	MT AA	All possible AA + Stop codon	NA
WT_Mut	WT and MT	Concatenation of WT and MT AAs	All 420 possible AA combinations	NA
AA1_polarity	WT polarity	Polarity of AA1 side chain	hydrophobic, special, uncharged,+,-	http://www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/abbreviation.html#refs
AA2_polarity	MT polarity	Polarity of AA2 side chain	hydrophobic, special, uncharged,+,-	http://www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/abbreviation.html#refs
AA1_PI	WT pI	Isoelectric point of AA1	3.22-9.74	http://www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/abbreviation.html#refs
AA2_PI	MT pI	Isoelectric point of AA2	3.22-9.74	http://www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/abbreviation.html#refs
deltaPI	pI change	Difference between WT and MT pI values	(-6.52)-6.52	NA
Grantham	Grantham	Physicochemical distance between WT and MT AA	0-215	Grantham, R. *Science* (1974)
AA2_weight	WT weight	Molecular mass (Da)	75-204	http://www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/abbreviation.html#refs
AA1_weight	MT weight	Molecular mass (Da)	75-204	http://www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/abbreviation.html#refs
deltaWeight	Weight change	Difference between WT and MT weights	(-192)-192	NA
AA1vol	WT volume	AA1 volume ($Å^3$)	60.1-227.8	Zamyatnin, A.A. Prog. Biophys. Mol. Biol (1972)
AA2vol	MT volume	AA2 volume ($Å^3$)	60.1-227.8	Zamyatnin, A.A. Prog. Biophys. Mol. Biol (1972)
deltavolume	Volume change	Difference between WT and MT volumes	(-167.7)-167.7)	NA
B_factor	B factor	B/Temperature factor from X-ray crystallography	0-84.35	Kabsch, W. & Sander, C. (1983)
Accessibility	Solvent accessibility	Number of water molecules in contact with this residue *10	0-238	Kabsch, W. & Sander, C. (1983)
dssp_sec_str	Secondary structure	Secondary structure	B, E, G, H, S, T, None	Kabsch, W. & Sander, C. (1983)
aa1_psic	WT likelihood	AA1 log likelihood ratio	(-4.083)-(-0.596)	Adzhubei et al. 2010
aa2_psic	MT likelihood	AA2 log likelihood ratio	-5.621-(-0.807)	Adzhubei et al. 2010
delta_psic	Likelihood change	Change in log likelihood ratios	-3.07 - 4.868	Adzhubei et al. 2010
phi_psi_reg	Phi-psi	Region of the Ramachandran map	A, B, I, L, None	Adzhubei et al. 2010
delta_solvent_accessibility	Accessibility change	Predicted change in solvent accessibility	0 - 2.92	Adzhubei et al. 2010
mut_msa_congruency	MSA Substitution score	maximum homology of the AA2 to all sequences in multiple alignment	0.044 - 47.42	Adzhubei et al. 2010
mut_mut_msa_congruency	MT MSA Substitution	ximum homology of the AA2 to the sequences in multiple alignment with the mutant resic	1.462 - 47.42	Adzhubei et al. 2010
seq_ind_closest_mut	Homolog with MT	Query sequence identity with the closest homologue deviating from the AA1	9.03 - 93.7	Adzhubei et al. 2010
evolutionary_coupling_avg	Evolutionary coupling	Mean evolutionary coupling score	0-0.11	derived from Hopf, et al 2017 evo couplings scores

Abbreviations: WT = wild-type, AA, amino acid, MT = mutant, H = α-helix B = residue in isolated β-bridge, E = extended strand, participates in β ladder, G = 3-helix (310 helix), T = hydrogen bonded turn, S = bend

Appendix B Table 3 Grid search values for hyperparameter tuning and final hyperparameter values used to train Envision.

Tuning round	Hyperparameter	Tested values	Optimum
1	Maximum number of decision trees	10, 25, 50, 100, 250	50
2	Maximum tree depth	2, 6, 10, 25, 50	6
	Minimum number of observations in terminal node of decision tree	2, 6, 10, 25, 50	50
3	Loss reduction required to add another branch to decision tree	0, 0.1, 0.2, 0.3, 0.4, 0.5	0.5
4	Feature subsample proportion at each iteration	0.6, 0.7, 0.8, 0.9	0.6
	Variant effect score subsample proportion at each iteration	0.6, 0.7, 0.8, 0.9	0.9
5	Increase iteration # 5-fold and reduce learning rate from 0.1 to 0.01 to compensate.	Trees = 250; Shrinkage = 0.01	

Appendix B Table 4 Importance of each feature in Envision's gradient boosted model.

Feature	Importance	Type
B factor	1347	Structural
Solvent accessibility	1299	Structural
Homolog with MT	1025	Evolutionary
WT likelihood	897	Evolutionary
Evolutionary coupling	839	Evolutionary
Likelihood change	628	Evolutionary
Accessibility change	536	Structural
MT likelihood	477	Evolutionary
MSA Substitution score	341	Evolutionary
Proline mutant	314	Physicochemical
Grantham	312	Physicochemical
WT weight	279	Physicochemical
Volume change	244	Physicochemical
WT volume	230	Physicochemical
WT pI	230	Physicochemical
Weight change	190	Physicochemical
MT weight	156	Physicochemical
MT volume	133	Physicochemical
Cysteine mutant	106	Physicochemical
MT pI	101	Physicochemical
Helix structure	99	Structural
pI change	93	Physicochemical
Beta strand structure	92	Structural
MT polarity	91	Physicochemical
WT polarity	77	Physicochemical

*Importance was determined by counting the number of times each feature occurred in the Envision decision tree ensemble.

VITA

Vanessa E. Gray was born in Chandler, Arizona (USA) on April 30, 1989. Between 2003 and 2007 she studied genetics at Arizona State University in Tempe. She received a graduate research fellowship from the National Science Foundation to study in Prof. Douglas M. Fowler's group at the University of Washington from 2012 to 2017.