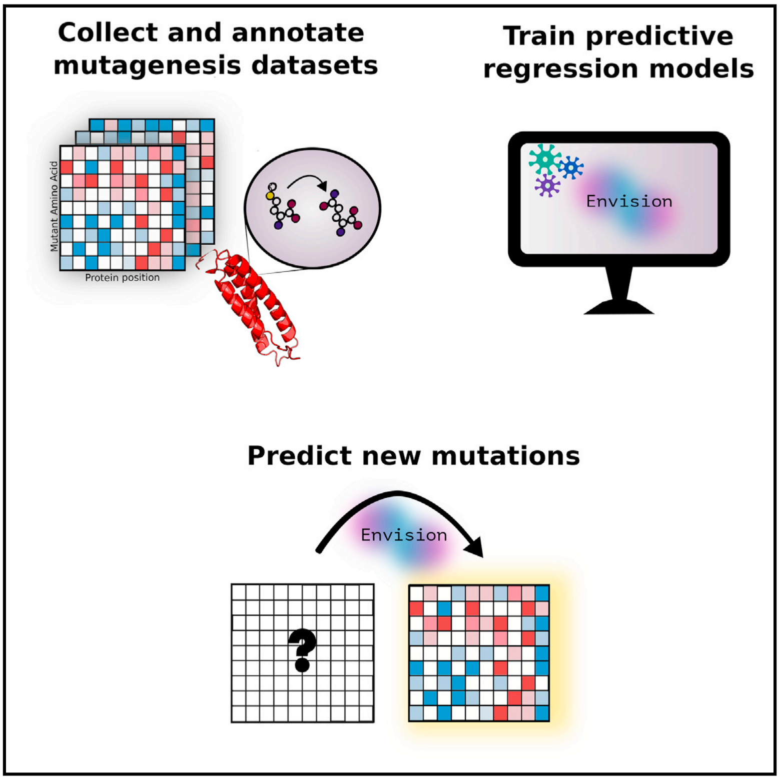
Cell Systems

Report

Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data

Graphical Abstract


Collect and annotate mutagenesis datasets
Train predictive regression models
Mutant Amino Acid
Protein position
Envision
Predict new mutations
Envision
?

Highlights

- Large-scale, quantitative mutagenesis data offer a novel source of training data
- Envision outperforms other missense variant effect predictors on independent data
- More mutagenesis data will improve Envision's predictive performance
- Envision predictions are available for download: https://envision.gs.washington.edu

Authors

Vanessa E. Gray, Ronald J. Hause, Jens Luebeck, Jay Shendure, Douglas M. Fowler

Correspondence

dfowler@uw.edu

In Brief

We present Envision, an accurate predictor of protein variant molecular effect, trained using large-scale experimental mutagenesis data.

Gray et al., 2018, Cell Systems 6, 116–124
January 24, 2018 © 2017 Elsevier Inc.
https://doi.org/10.1016/j.cels.2017.11.003

Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data

Vanessa E. Gray,[1] Ronald J. Hause,[1] Jens Luebeck,[1] Jay Shendure,[1,2] and Douglas M. Fowler[1,3,4,*]
[1]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
[2]Howard Hughes Medical Institute, Seattle, WA 98195, USA
[3]Department of Bioengineering, University of Washington, Seattle, WA 98195, USA
[4]Lead Contact
*Correspondence: dfowler@uw.edu
https://doi.org/10.1016/j.cels.2017.11.003

SUMMARY

Large datasets describing the quantitative effects of mutations on protein function are becoming increasingly available. Here, we leverage these datasets to develop Envision, which predicts the magnitude of a missense variant's molecular effect. Envision combines 21,026 variant effect measurements from nine large-scale experimental mutagenesis datasets, a hitherto untapped training resource, with a supervised, stochastic gradient boosting learning algorithm. Envision outperforms other missense variant effect predictors both on large-scale mutagenesis data and on an independent test dataset comprising 2,312 TP53 variants whose effects were measured using a low-throughput approach. This dataset was never used for hyperparameter tuning or model training and thus serves as an independent validation set. Envision prediction accuracy is also more consistent across amino acids than other predictors. Finally, we demonstrate that Envision's performance improves as more large-scale mutagenesis data are incorporated. We precompute Envision predictions for every possible single amino acid variant in human, mouse, frog, zebrafish, fruit fly, worm, and yeast proteomes (https://envision.gs.washington.edu/).

INTRODUCTION

Mutations have the power to 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 (Vigneri et al., 2010). 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 (Karczewski et al., 2017; Zou et al., 2016). This example highlights how 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 (MacArthur et al., 2014).

Experiments can reveal a variant's molecular effect, and recent advances in multiplex assays have enabled the assessment of large numbers of variants (Fowler and Fields, 2014; Gasperini et al., 2016). 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. Thus, variant effect predictors such as PolyPhen2 (Adzhubei et al., 2010), SIFT (Sim et al., 2012), SNAP2 (Hecht et al., 2015), Evolutionary Action (Katsonis and Lichtarge, 2014), CADD (Kircher et al., 2014), and a host of others (Tang and Thomas, 2016) 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, and is thus optimized to predict clinical variant effects (Adzhubei et al., 2010). SNAP2, an ensemble of neural network models, is trained on human pathogenic and neutral variants as well as variants that affect molecular function (Hecht et al., 2015). 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 (Katsonis and Lichtarge, 2014), 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 (Hopf et al., 2017). 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 versus 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 (Stenson et al., 2012), Online Mendelian Inheritance of Man (Amberger et al., 2015), and ClinVar (Landrum et al., 2013) all categorize variants as clinically deleterious or benign. Swiss-Prot and the Protein Mutant Database contain categorical measures of variant effects in laboratory

© 2017 Elsevier Inc.

assays. Second, most predictors focus on predicting the clinical effect of human variants rather than the molecular effects on protein function (Adzhubei et al., 2010; Sim et al., 2012). 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 (Rodriguez-Viciana et al., 2006; Wan et al., 2004). Variants of PTEN variants can drive carcinogenesis when they occur somatically, or can cause autism or a tumor syndrome when they occur in the germline (Mester and Eng, 2013). Thus, we 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, we 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 (Fowler and Fields, 2014; Fowler et al., 2014). Sequencing reveals how each variant's frequency changes during selection, yielding 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 such as 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 a protein. We hypothesize that large-scale mutagenesis datasets contain informative and generalizable patterns that can be used to predict variant effects in disparate proteins.

Here, we 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 (Friedman, 2002). To maximize Envision's generalizability, proteins in the Envision training set have disparate structures and functions, and are drawn from diverse organisms. We 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 with the observed variant effects for the excluded protein. We also assess performance using independent variant effect data that were not generated by deep mutational scanning nor included in Envision's training. Envision's predictions are generally more accurate than other state-of-the-art predictors. Envision's prediction accuracy is also consistent across different amino acids, unlike other predictors that perform well on some amino acids and poorly on others. We precomputed Envision predictions for all possible single amino acid variants of proteins in the human, mouse, fruit fly, clawed frog, zebrafish, worm, and yeast proteomes. We provide a web-based tool allowing users to visualize and explore predicted protein sequence-function maps, which can be used to prioritize variants. Envision is available at https://envision.gs.washington.edu.

RESULTS

Data Collection and Curation

We collected previously published, large-scale mutagenesis datasets with quantitative measures of variant effect on protein function. Exploratory analysis led to the following inclusion criteria: (1) the experiment must have measured single amino acid variant effects, rather than averaging across different genetic backgrounds; (2) the experiment must have been on a natural protein instead of a designed protein; and (3) the experiment must have quantitated effects for at least ~50% of all possible variants of the mutagenized region. Ultimately, deep mutational scans of ten proteins from 12 studies comprising 28,545 single amino acid variant effects met these criteria (Figure 1A and Table S1). Variant coverage ranged from ~50% for the Ube4b domain of murine E3 ligase to 100% for the immunoglobulin G-binding domain of influenza protein G and the PDZ domain of human PSD-95 (Figure 1B). Variant coverage depended on experimental details such as the protocol used for library generation (e.g., doped oligomer [Matteucci and Heyneker, 1983] versus site saturation mutagenesis [Jain and Varadarajan, 2014]), the number of clones generated, and the sequencing depth. The proteins in the dataset were distinct, coming from different organisms, having different structures, and having functions ranging from catalysis to peptide binding (Table S1). To make datasets comparable, we normalized variant effect scores in each dataset such that variants that were more active than wild-type had a variant effect score greater than 1, wild-type-like variants had a score of 1, and variants that were less active than wild-type had a score less than 1 (Figures S1 and 1C).

Next, we annotated each variant with 27 biological, structural, and physicochemical features. The biological features captured evolutionary constraints using both site-specific and co-varying conservation metrics. The structural features include local density and solvent accessibility, 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 1D and Table S2).

Predicting Quantitative Variant Effects

We first tested whether a stochastic gradient boosting regression algorithm could model the relationships between our 27 features and quantitative variant effect scores for each protein individually. To train each single-protein model, we tuned hyperparameters, such as the number of decision trees in the ensemble and tree depth, using 10-fold cross-validation. After hyperparameter tuning we reserved 20% of mutations for testing, allowing us to estimate the generality of each model to unseen variants. Nine of the 12 models performed well (median Pearson's R = 0.83, Spearman's $\rho = 0.80$, Figure 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, $\rho = 0.35$).

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 (Table S1). We hypothesized

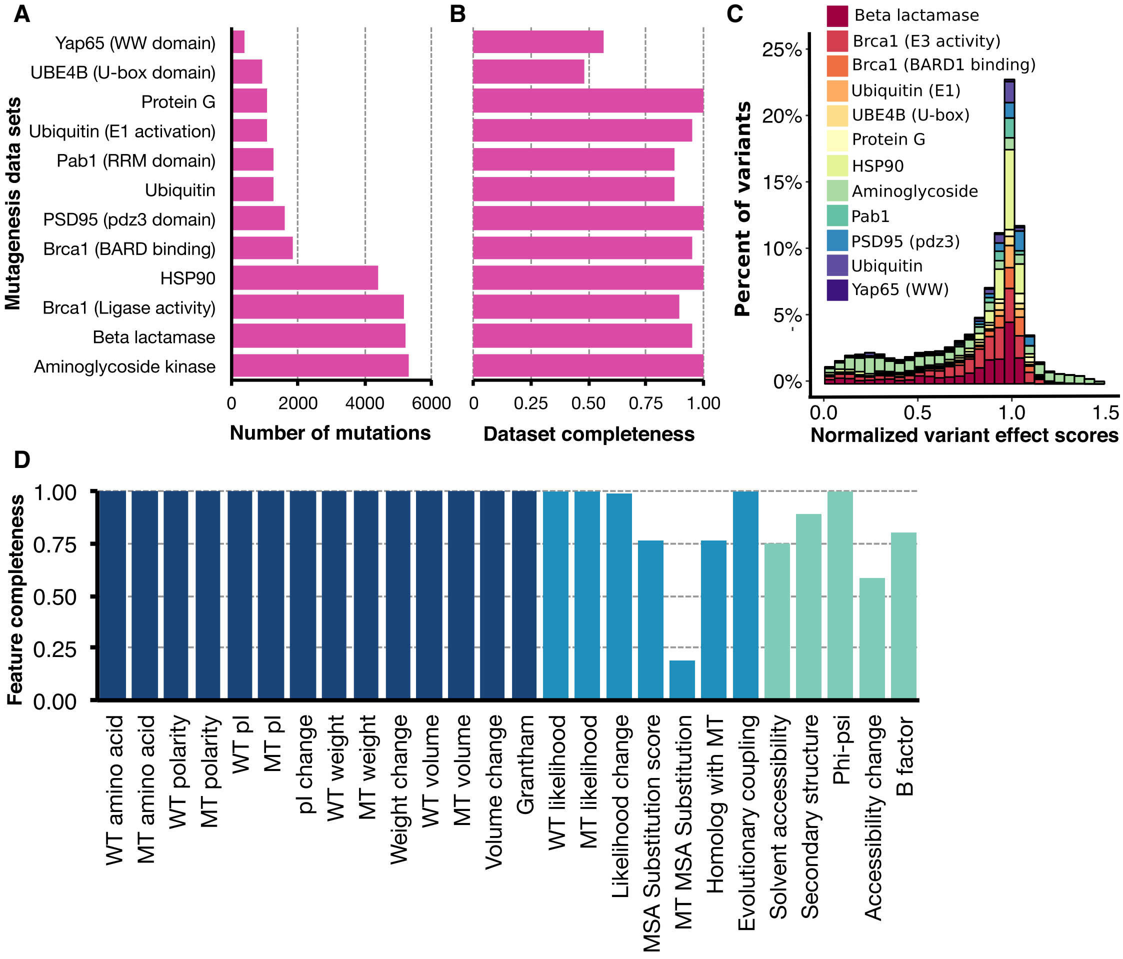


Figure 1. Large-Scale Mutagenesis Data and Descriptive Features Used to Train Envision

(A and B) 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 of 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 (see Figure S1 for non-normalized score distribution). Each collected variant was annotated with 27 features, which describe physicochemical (dark blue), evolutionary (blue), or structural (green) variant attributes (Table S2).

(D) The proportion of variants in the collected large-scale mutagenesis datasets having each feature is shown (WT, wild-type; MT, mutant).

that poor performance arose because correlations between the features and variant effect scores were low (Figure S2). Low correlation might occur because the assays did not test every function of these proteins. For instance, BCRA1 RING domain variants were assayed for E3 ligase activity and BARD binding. However, BRCA1 has many functions and interacts with >25 other proteins (Deng and Brodie, 2000; Kerrien et al., 2012). Another possibility is that these two datasets were missing some structural features. However, the YAP65 WW domain dataset, missing the same features, resulted in an accurate model. Thus, we could not identify the cause of poor performance in the BRCA1 RING domain and E4B ubiquitin ligase models. We excluded these three datasets from subsequent analyses.

For most proteins, our feature set and learning procedure generated accurate models of variant effect. Beyond validating our approach, these single-protein models enabled us to complete each large-scale mutagenesis dataset by predicting missing variant effect scores (Table S3). For example, we used the Pab1 model (R = 0.86; ρ = 0.79) to predict the ~20% of scores that were missing, completing the Pab1 dataset (Figure 2B).

Next, we trained a global model with the 21,026 empirically derived variant effect scores in the nine large-scale mutagenesis datasets. We tuned hyperparameters using a leave-one-protein-out (LOPO) approach designed to avoid protein-specific overtraining (Figure S3 and Table S4). Once hyperparameters were tuned, we trained Envision with all available data, except for a random 5% of variant effect scores that we 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 (Figure S4). Envision predicted the training data well (R = 0.79, ρ = 0.76; Figure 3A).

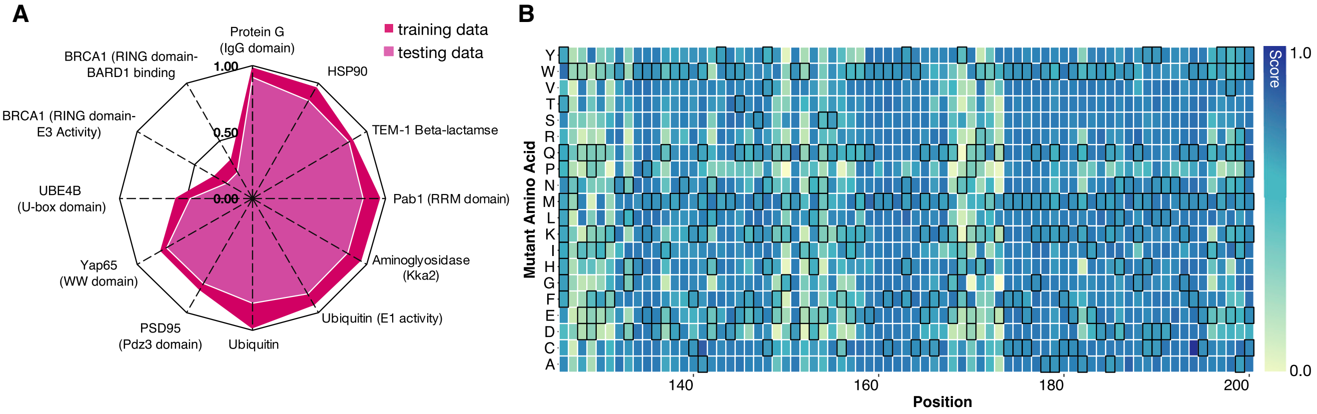


Figure 2. Protein-Specific Gradient Boosting Models Can Accurately Predict Variant Effect Scores

We 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 yellow. Predicted effects are denoted by black borders.

Assessing Envision's Performance

To evaluate Envision's performance, we employed a jack-knife LOPO approach. Here, we repeated the training procedure described above, leaving one protein completely out of the hyperparameter tuning and model training process. We then used the resulting model to predict variant effect scores for the left-out protein and determined performance. We 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 (Figures 3B and S5). To determine the effect of our variant effect score normalization scheme on model training and performance, we compared LOPO models trained using either normalized or non-normalized variant effect scores. 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 versus 0.39, median ρ = 0.51 versus 0.35; Figure S6). This result highlights the utility of our normalization scheme.

Next, we compared our LOPO models' performance with that of other predictors. PolyPhen2 is trained to predict the categorical clinical effect of variants, but also generates a numerical score. This score is the naive 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 versus 0.17; ρ = 0.36 versus 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 versus 0.03; ρ = 0.36 versus 0.04). PolyPhen2 and SIFT were not designed to predict variant effect magnitude, and our results confirm that they should not be used to do so.

SNAP2, EVmutation, and Evolutionary Action were developed to predict variant effect magnitude (Hecht et al., 2015; Hopf et al., 2017; Katsonis and Lichtarge, 2014). Evolutionary Action scores could not be obtained by batch query, preventing us from including them in our analysis. 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 versus 0.44; Figure 3C). EVmutation predicts variant effect magnitude using either an epistatic or an independent conservation-based unsupervised statistical model. Our LOPO models outperformed EVmutation's epistatic model on six out of nine datasets (median R 0.56 versus 0.47) and EVmutation's independent model on seven of nine (mean R 0.56 versus 0.48; Figure 3C). An equivalent analysis using Spearman's ρ revealed similar results (Figure S7). 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, respectively.

Next, we analyzed which factors led to our improved performance. Envision's features are similar to those used by SNAP2 and PolyPhen2, so the improvement we observed is not likely due to feature choice. Instead, we 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 using deep mutational scanning data. However, we quantified the impact of our cross-validation approach by comparing the performance of models trained using standard 10-fold cross-validation with models trained using our LOPO scheme. We found our LOPO approach improved performance by ~10%–20% over all protein datasets compared with 10-fold cross-validation (median R = 0.56 versus 0.45, ρ = 0.50 versus 0.45; Figure S8). We

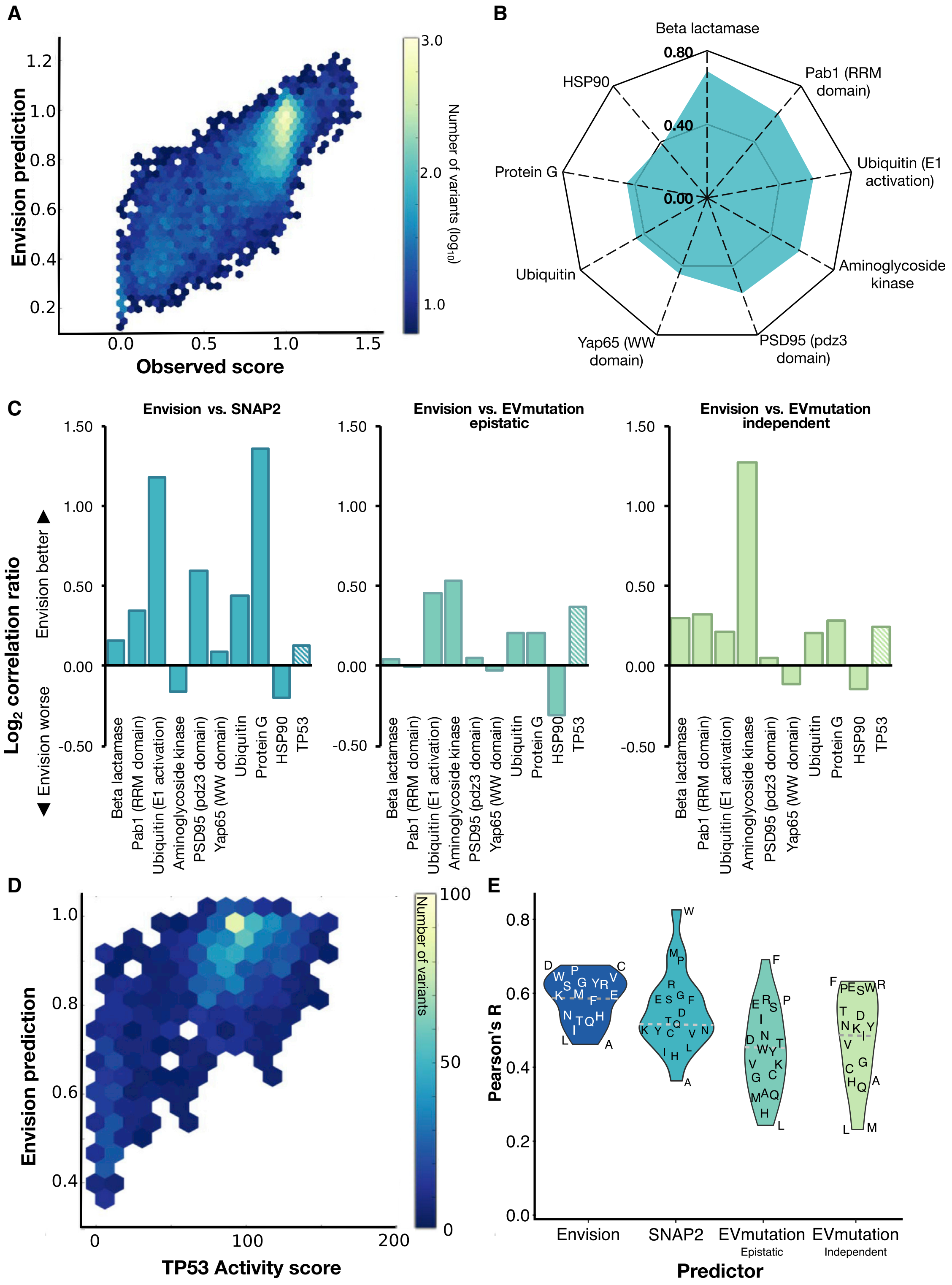


Figure 3. Envision Outperforms Other Quantitative Variant Effect Predictors

(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).

(B) To evaluate performance on data not used in training, we retrained models excluding each one of the nine proteins (see Figures S3 and S4 for cross-validation scheme and training performance). A radar plot shows the correlation (Pearson's R) between predicted and observed variant effect scores when the indicated protein was left out (see Figure S5 for scatterplots).

(legend continued on next page)

suggest that our LOPO approach, designed to yield generalizable models, was especially important given that our training dataset contained relatively few proteins.

Our LOPO analysis demonstrated that Envision provided improved quantitative predictions of variant effects measured using deep mutational scanning. However, Envision's performance advantage might have arisen because it learned deep mutational scanning-specific patterns in the data. To ensure that Envision was not overfitted to deep mutational scanning data, we obtained a TP53 tumor suppressor mutagenesis dataset whereby the effects of 2,312 variants on TP53 transactivation were measured individually using a fluorescent reporter (Kato et al., 2003). We predicted these TP53 variant effect scores using Envision, which was trained on all nine large-scale mutagenesis datasets. The TP53 data were never used, directly or indirectly, in 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 ($R = 0.58$, $\rho = 0.53$; Figures 3C and 3D). Importantly, Envision outperformed SNAP2 ($R = 0.53$; $\rho = 0.50$), whose training dataset included the effects of ~400 human TP53 mutations, and EVmutation (epistatic $R = 0.45$, $\rho = 0.49$; independent $R = 0.49$, $\rho = 0.52$; Figure 3C). Thus, Envision learned patterns of the molecular effects of variants that do not depend on the measurement method.

Next, we sought to determine whether Envision performance depended on the identity of either the mutant or wild-type amino acid. We evaluated performance on the TP53 dataset to enable comparison with EVmutation and SNAP2. We found that Envision prediction performance did not depend much on the identity of the mutant amino acid (Figures 3E and S9A). 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 were also apparent for the wild-type amino acid, where EVmutation predicted mutations from wild-type cysteine well ($R = 0.82$, $\rho = 0.71$) and wild-type aspartic acid poorly ($R = 0.02$, $\rho = 0.05$; Figure S9B). Consequently, in addition to greater overall accuracy, Envision performance was more consistent.

Finally, we assessed the utility of Envision scores for clinical effect prediction, evaluating performance by constructing receiver-operating characteristic (ROC) curves using variants annotated as either pathogenic or benign in ClinVar. Envision predictions were better than random guessing (area under ROC curve [AUROC] = 0.72), but not as good as PolyPhen2, CADD, and SIFT (AUROC = 0.86, 0.85, and 0.84, respectively; Figure S10). 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 data (Grimm et al., 2015). 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, we note that the rate at which training datasets grow in the coming years may be much greater for deep mutational scans than for clinical variants.

Feature Importance and Future Improvements

Features known to be predictive of variant effects, including solvent accessibility and evolutionary conservation, were the most highly represented in the Envision decision tree ensemble (Figure 4A and Table S5) (Kumar et al., 2009; Saunders and Baker, 2002). However, unlike for other feature-driven predictors (Adzhubei et al., 2010; Hecht et al., 2015), we found that the mutant amino acid identity was informative. This amino acid identity effect was largely driven by proline. Proline variants are generally disruptive of protein function and, indeed, proline variants were the most damaging substitutions in the large-scale mutagenesis datasets (proline mean effect score = 0.60 versus all-amino-acid mean = 0.81; paired t test, $p << 0.001$, $n = 8$; Figure S11). Envision predicted the effects of proline variants about as accurately as the effects of other variants (Figure S12). Thus, rather than simply predicting that all proline variants were strongly damaging, Envision predicted the degree to which proline variants maintain or disrupt function.

Structural and evolutionary features are important for Envision's predictions but are not always available. Thus, we quantified predictive performance when these features were missing by masking them for each of the nine LOPO models when they predicted the left-out protein's variant effect scores. As expected, models performed worse without structural features. For example, the β-lactamase LOPO model predicted β-lactamase variant effect scores 15% worse when structural features were masked ($R = 0.69$ versus 0.59; Figure S13). Similarly, the β-lactamase LOPO model predicted β-lactamase variant effect scores 13% worse when evolutionary features were masked ($R = 0.69$ versus 0.60). Across the nine LOPO models, we found that masking structural features degraded performance by 39% and masking evolutionary features degraded performance by 18%. Thus, we strongly encourage users to consider feature completeness when using Envision's predictions. Feature information is available, along with predictions, on the Envision website. We note that all feature-driven predictors suffer when key features are unavailable.

Finally, we determined how the number of proteins in our training dataset affected Envision performance. We trained

(C) We also compared the leave-one-protein-out (LOPO) models with SNAP2 (left panel), EVmutation-epistatic (middle panel), and EVmutation-independent (right panel). The $\log_2$ ratio of each LOPO model's Pearson's R to another predictor Pearson's R on the left-out data is shown. Hashed bars on the right indicate relative performance on a set of 2,312 TP53 transactivation activity scores measured in a low-throughput assay and not used in training (see Figure S7 for raw comparison).

(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 (see Figures S9A and S9B for heatmap of correlations).

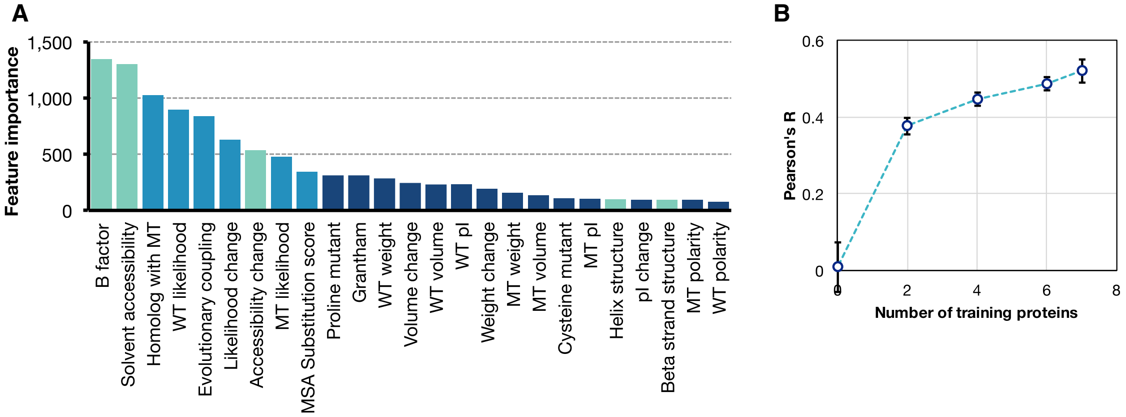


Figure 4. Envision Is an Interpretable Model that Will Improve with More Training Data

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) feature is shown (WT, wild-type; MT, mutant). See Figures S11 and S12 for proline feature analysis.

(B) To assess the impact of adding more training data to Envision, we conducted a downsampling analysis. Models were trained with increasing numbers of randomly selected protein datasets and tested on mutations from proteins withheld from training. 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 SD of correlation coefficients obtained from ten random samplings of proteins to include in the training set. A naive 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 naive model indicate the SD of correlation coefficients obtained from ten different feature randomizations. See Figure S13 for left-out feature analysis.

versions of Envision with different numbers of proteins and tested it on the left-out proteins. We found that model performance increased as more proteins were used in training, suggesting that accumulation of more data will improve Envision's predictive performance (Figure 4B).

Availability of Envision Predictions

Envision predictions are available for proteins from seven commonly studied organisms: human (n = 20,130), mouse (n = 16,836), fruit fly (n = 3,375), clawed frog (n = 1,704), zebrafish (n = 2,982), worm (n = 3,802), and yeast (n = 8,322). We provide predictions for all 19 alternative amino acids at each position, with batch query and download options available. Along with predictions, features are also available for download.

DISCUSSION

We developed Envision, the first variant molecular effect predictor trained on large-scale mutagenesis data. Envision accurately predicts variant effects in large-scale mutagenesis data withheld from training as well as variant effects from low-throughput experiments. Overall, Envision outperforms other quantitative predictors such as 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. As more large-scale mutagenesis data become available, Envision will continue to improve.

Envision also has limitations. Envision's predictions are provided as quantitative scores that range from ~0 to ~1, where scores less than 1 are damaging in comparison with wild-type. Envision can predict the scores of strongly damaging and wild-type-like mutations well, but predicts mutations of intermediate effect less well (Figure S5). Envision also relies on structural and evolutionary features that are not available for every protein, and predictive performance degraded when these features were missing. Thus, while Envision predictions are available for millions of variants, we recommend caution when key features are missing. The Envision web tool reveals missing features for each prediction.

To train Envision, we employed large-scale mutagenesis data from two types of deep mutational scans. One type is based on a generalized selection for protein function whereas the other is based on selection for a specific protein function. Specific selections could fail to capture the effect of variants on other functions of the protein, such as binding to a different substrate or catalysis. Envision was trained using data from both generalized and specific deep mutational scans and did not distinguish between them. Therefore, Envision predicts generalized variant effects and does not distinguish between specific molecular effects such as enzymatic activity for one substrate or another, or binding versus catalysis. Collection of more large-scale mutagenesis data for the specific molecular effects of variants may enable the development of predictors that capture these specific functional effects.

We anticipate that Envision will be useful for identifying candidate variants that tune protein activity levels. Envision's predictions of molecular effect may also be useful when the relationship between protein function and disease is clear. Furthermore, Envision will continue to improve as new datasets become available.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
  - ○ Training Data Collection
  - ○ Normalization
  - ○ Variant Annotation
  - ○ Machine Learning
  - ○ Single Protein Models
  - ○ Training Envision
  - ○ Leave-one-protein-out (LOPO) Models
  - ○ Downsampling Analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes 13 figures and 5 tables and can be found with this article online at https://doi.org/10.1016/j.cels.2017.11.003.

AUTHOR CONTRIBUTIONS

The project was conceived and designed by V.E.G., R.J.H., J.L., J.S., and D.M.F. V.E.G. and R.J.H. performed data analysis and generated figures. All authors wrote the manuscript.

ACKNOWLEDGMENTS

We thank Bill Noble and Christine Queitsch for insightful comments. This research was supported by research grants from the National Science Foundation to V.E.G. (DGE-1256082), and the NIH to J.S. (DP1HG007811) and D.M.F. (R01GM109110). R.J.H. is a Damon Runyon Fellow supported by the Damon Runyon Cancer Research Foundation (DRG-2224-15). J.S. is an Investigator of the Howard Hughes Medical Institute.

Received: May 20, 2017
Revised: August 30, 2017
Accepted: November 3, 2017
Published: December 6, 2017

REFERENCES

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.

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. Nucleic Acids Res. *43*, 789–798.

Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference (ACM), pp. 785–794.

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

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

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.

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

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

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

Grimm, D.G., Azencott, C.-A., Aicheler, F., Gieraths, U., MacArthur, D.G., Samocha, K.E., Cooper, D.N., Stenson, P.D., Daly, M.J., Smoller, J.W., et al. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum. Mutat. *36*, 513–523.

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

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.

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.

Jones, E., Oliphant, E., and Peterson, P. (2001). SciPy: Open Source Scientific Tools for Python. http://www.scipy.org/.

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

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

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. USA *100*, 8424–8429.

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.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., et al. (2012). The IntAct molecular interaction database in 2012. Nucleic Acids Res. *40*, 841–846.

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.

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.

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. Nucleic Acids Res. *44*, 862–868.

Low, Y., Gonzalez, J., Kyrola, A., Bickson, D., Guestrin, C., and Hellerstein, J.M. (2012). Distributed GraphLab: a framework for machine learning in the cloud. arXiv 1204.6078.

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.

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. Nucleic Acids Res. *11*, 3113–3121.

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

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

Pedregosa, F., Thirion, B., Michel, V., Gramfort, A., and Varoquaux, G.L. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

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.

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. Nucleic Acids Res. *39*, 392–401.

Saunders, C.T., and Baker, D. (2002). Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J. Mol. Biol. *322*, 891–901.

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. Nucleic Acids Res. *40*, 452–457.

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.

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.

Tang, H., and Thomas, P.D. (2016). Tools for predicting the functional impact of nonsynonymous genetic variation. Genetics *203*, 635–647.

van der Walt, S., Colbert, S.C., and Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. Comput. Sci. Eng. *13*, https://doi.org/10.1109/MCSE.2011.37.

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

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., and Marais, R. (2004). Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. Cell *116*, 855–867.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer).

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.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Python/2.7.3	Python	https://www.python.org/
Numpy/1.13.1	van der Walt et al. (2011)	http://www.numpy.org/
GraphLab/2.1	Low et al. (2012)	https://turi.com/
Scipy/0.19.1	Jones et al. (2001)	https://www.scipy.org/
scikit-learn/0.17.0	Pedregosa et al. (2011)	http://scikit-learn.org/stable/
R version 3.2.3	R	https://cran.r-project.org/
ggplot2/2.2.1	Wickham (2016)	http://ggplot2.org/
reshape2/1.4.2	Wickham (2016)	https://cran.r-project.org/web/packages/reshape2/index.html
DSSP	Kabsch and Sander, (1983)	http://swift.cmbi.ru.nl/gv/dssp/
Polyphen2 (annotations and predictions)	Adzhubei et al. (2010)	http://genetics.bwh.harvard.edu/pph2/dokuwiki/downloads
SIFT predictions	Ng and Henikoff (2001)	http://sift.jcvi.org/
EVmutation predictions	Hopf et al. (2017)	https://marks.hms.harvard.edu/evmutation/
SNAP2 predictions	Hecht et al. (2015)	https://rostlab.org/services/snap2web/
Envision	This study	https://github.com/FowlerLab/Envision2017
Other		
TEM1 β-lactamase	Firnberg	PubmedID: 24567513
Yap65 (WW domain)	Fowler	20711194
PSD95 (Pdz3 domain)	McLaughlin	23041932
Brca1 (RING domain)- E3 ligase activity	Starita	25823446
Brca1 (RING domain)- Bard1 binding	Starita	25823446
Aminoglycoside kinase	Melnikov	24914046
E4B (U-box domain)	Starita	23509263
Hsp90	Mishra	27068472
Ubiquitin	Roscoe	23376099
Pab1 (RRM domain)	Melamed	25671604
Ubiquitin - E1 activity	Roscoe	24862281
Protein G (IgG domain)	Olson	25455030

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Douglas M. Fowler (dfowler@uw.edu).

METHOD DETAILS

Training Data Collection

Published large-scale mutagenesis datasets were used as training data if they met several criteria. First, we 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, we only accepted mutational scans of native proteins assayed for native biological function. Third, we excluded scans in which the complete variant sequence was unknown. We also removed variants with more than one mutation. In total, we accepted twelve datasets comprising ~30,000 missense mutations. These scans were performed on proteins from different organisms: human, mouse, rat, *S. cerevisiae*, and bacteria (Table S1).

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, we normalized them. For each dataset, every variant effect score was normalized to the wild type score and then $\log_2$ transformed. Next, we 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 we instead subtracted 0 from each score in those datasets. Lastly, we 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 $S_{normalized} = (S_{reported\ i} - S_{median\ synonymous}) / (-S_{median\ bottom\ 1\%})+1$, 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.

Variant Annotation

Mutations were annotated with three general types of descriptive annotations: evolutionary, biochemical and structural (Table S2). Several evolutionary features used in our model were obtained using the PolyPhen2 annotation pipeline (Sunyaev et al., 1999). We also derived a measure of average mutational covariance between a given position and all other positions in a multiple sequence alignment from EVfold (Hopf et al., 2017). To obtain structural information, we use DSSP (http://www.cmbi.ru.nl/dssp.html) (Kabsch and Sander, 1983) and PDB files from the Protein Data Bank (http://www.rcsb.org/pdb/home/home.do) (Rose et al., 2011). Our biochemical annotations include measures of amino acid size, weight, volume, isoelectric point, and Grantham scores (Grantham, 1974).

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 (Friedman, 2002). We 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, we tuned six parameters in a stepwise fashion. First, we optimized for the number of decision trees in the ensemble. Next, we tuned the maximum depth of a decision tree and the minimum number of observances allowed in a terminal node of a tree. Then, we determined the value that the squared-loss must be reduced by in order to add an additional node to a tree. Finally, we identified the optimal proportion of variant effect scores and features used to train each tree. Once hyperparameters were tuned, we reduced the learning rate from 0.1 to 0.01 and increased the number of decision trees by five-fold. All tuned and trained models treat missing feature data as such, *i.e.*, no imputing procedures were performed. Instead, during training, the algorithm uses variants with missing features to determine how feature missingness should handled by the model at each tree node (Chen and Guestrin, 2016).

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, we 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 we 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.

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 (Figure S3). 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 (Table S4), we trained *Envision* with all available data except for a randomly selected 5% of which we excluded to evaluate model generality and ensure that the model was not overfitted. The resulting model was used to make all the Envision predictions available on our website.

Leave-one-protein-out (LOPO) Models

To estimate Envision's performance on proteins not used in model training, we 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. These LOPO models were used to estimate Envision's performance on proteins not included in the training set.

Downsampling Analysis

To evaluate the effect of additional training data on model performance, we trained models with 2, 4, 6 or 7 of the available nine protein datasets. Model training was performed as described in the Training Envision section above. Each model was used to predict variant effects in proteins that were not used during the training phase. Confidence intervals were generated by repeated rounds of randomly selecting proteins to use in the training phase (n = 8).

QUANTIFICATION AND STATISTICAL ANALYSIS

The details of the statistical test we conduct, as well as definitions of center and correlation can be found in the main text. Criteria for inclusion of deep mutational scanning data sets are described in the Method Details section of the STAR Methods.

DATA AND SOFTWARE AVAILABILITY

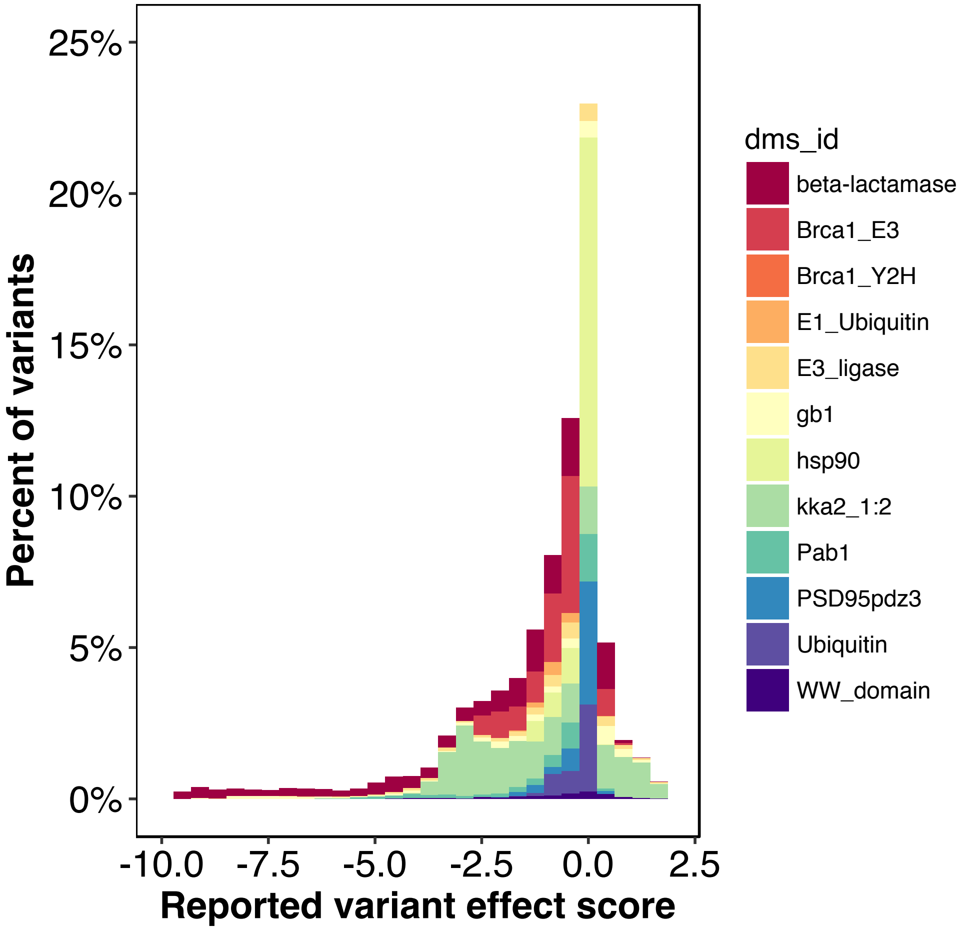
All data and software in this study are freely available. The training data set and all code used to train the models and generate the figures presented in this manuscript are available at https://github.com/FowlerLab/Envision2017. Envision predictions, along with feature annotations, are available at https://envision.gs.washington.edu/.

Cell Systems, Volume *6*

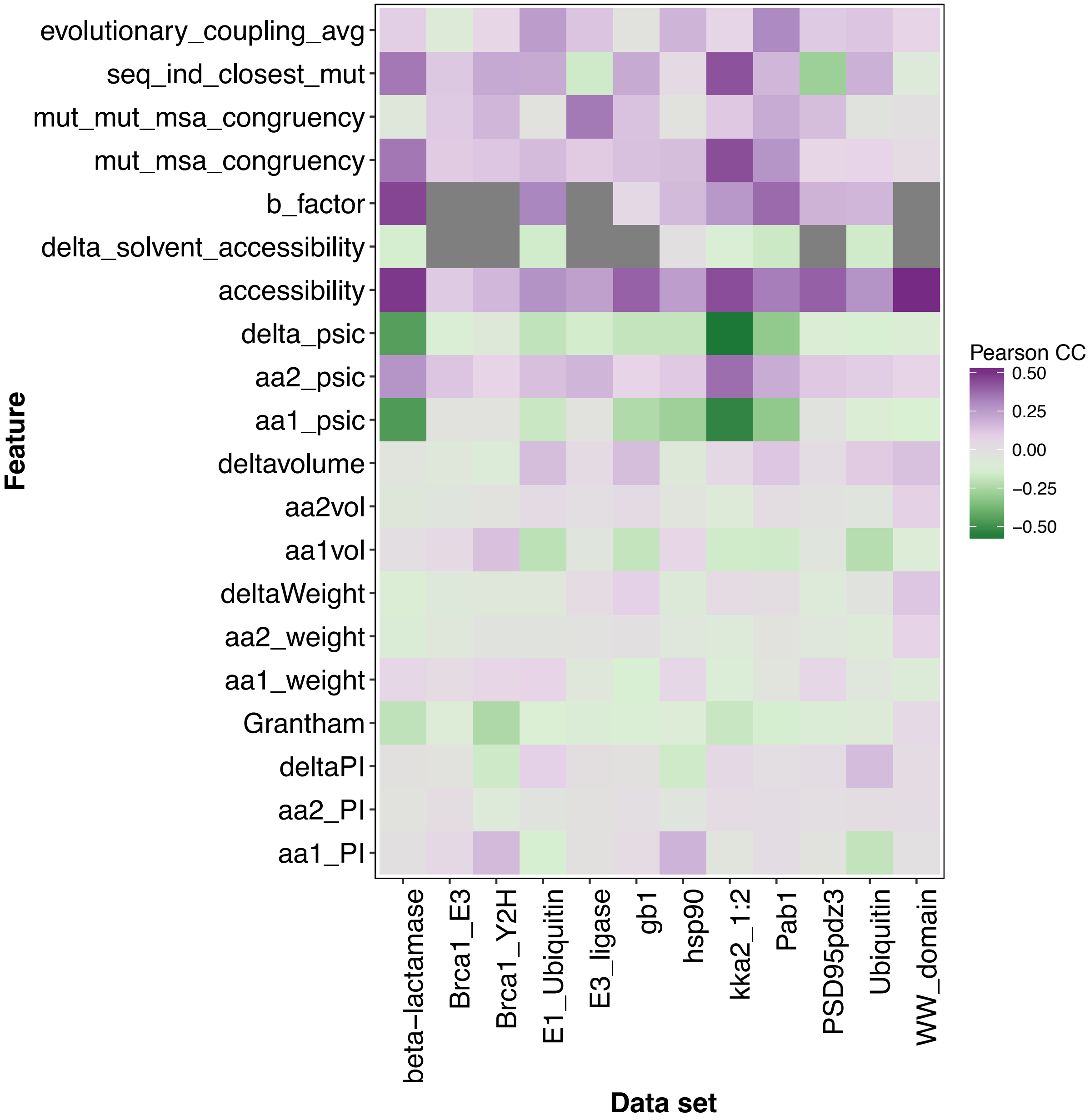
Supplemental Information

Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data

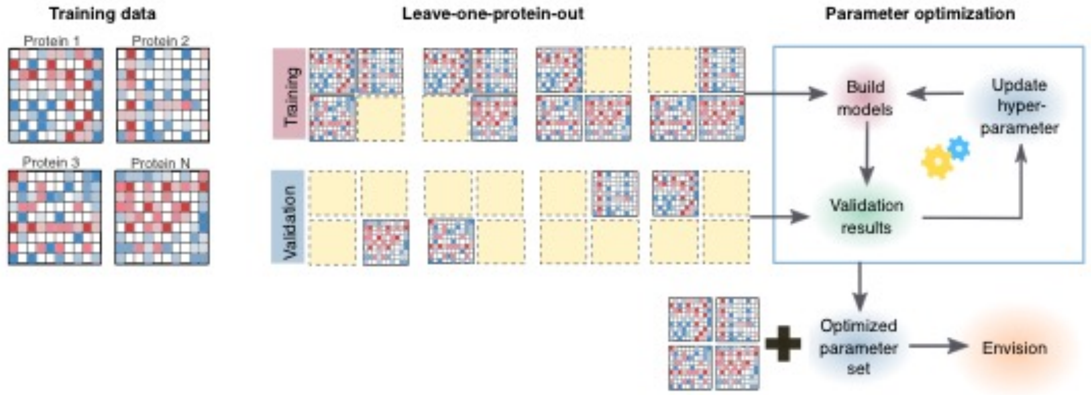
Vanessa E. Gray, Ronald J. Hause, Jens Luebeck, Jay Shendure, and Douglas M. Fowler



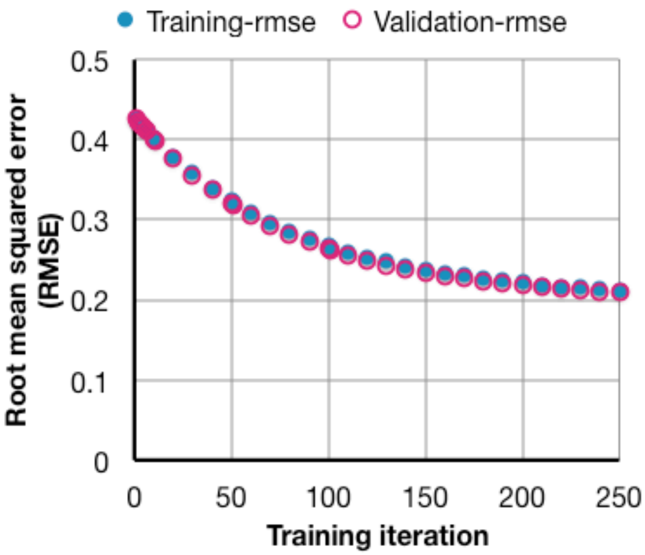
Supplementary Figure 1; related to Figure 1C. A histogram shows the distributions of reported variant effect scores from 12 large-scale mutagenesis data sets.



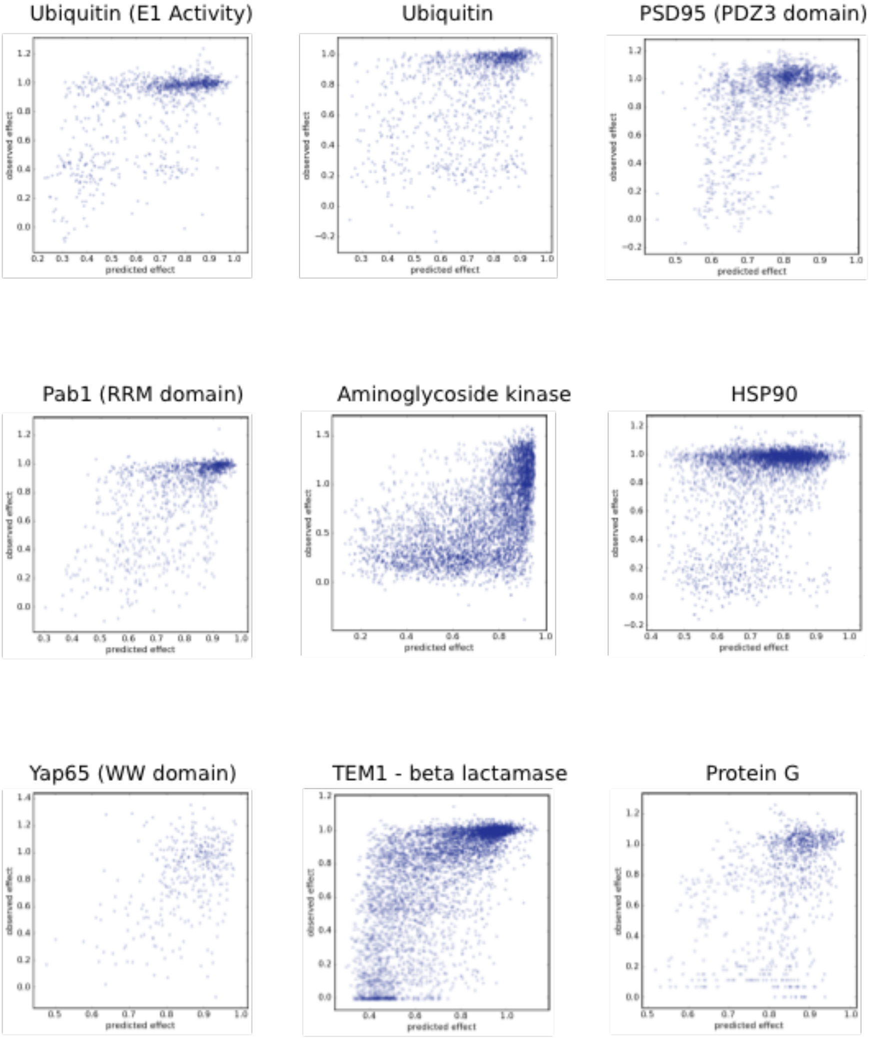
Supplementary Figure 2; related to figure 2A. A heatmap shows the Pearson correlation coefficient between descriptive feature values and variant effect scores for each large-scale mutagenesis data set. Note, E3 ligase, and BRCA1 datasets are missing B factor and predicted change in solvent accessibility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



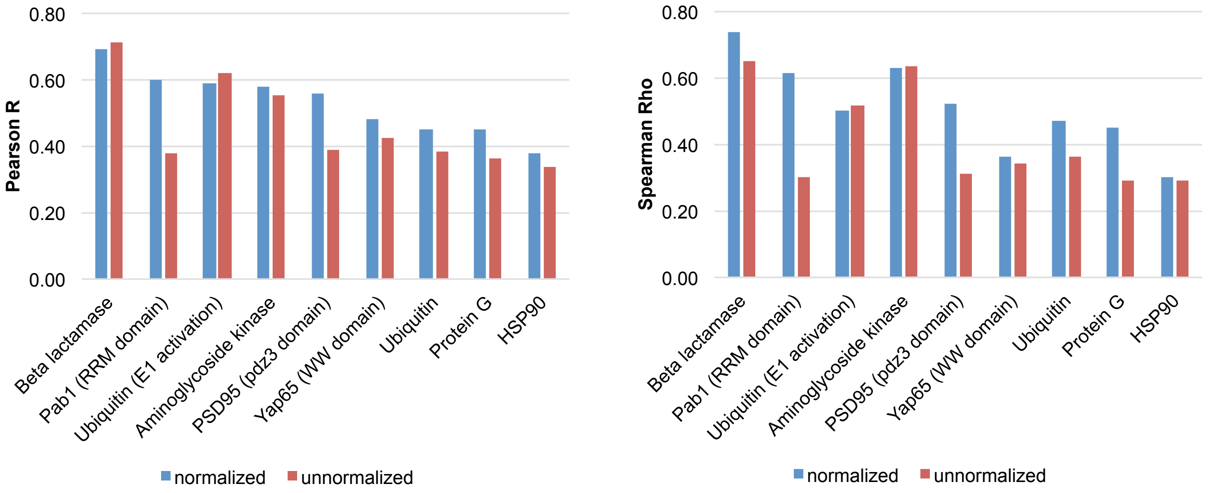
Supplementary Figure 3; related to Figure 3A. Our hyperparameter tuning scheme is designed to generate generalizable models. To determine the optimal values for each hyperparameter, we used a leave-one-protein-out cross-validation approach. To begin, we collected large-scale mutagenesis data sets and annotated them with features. Next, we 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, we test a set of hyperparameters using all testing and validation pair sets, and then update hyperparameters until all parameter values are evaluated. Once completed, we identify the parameter set that yields the most generalizable model, i.e., performs best on the left out protein's variant data set.



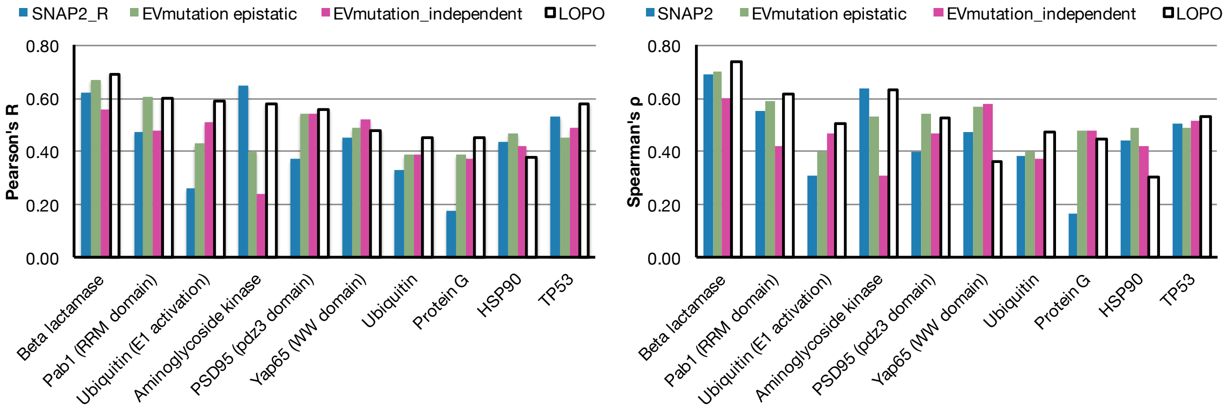
Supplementary Figure 4; related to Figure 3A. Training and testing data set RMSEs are very similar across iterations. While training Envision, 5% of data was withheld to determine the performance of the model as each tree was trained and added to the ensemble of decision trees. The plot shows the root 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.



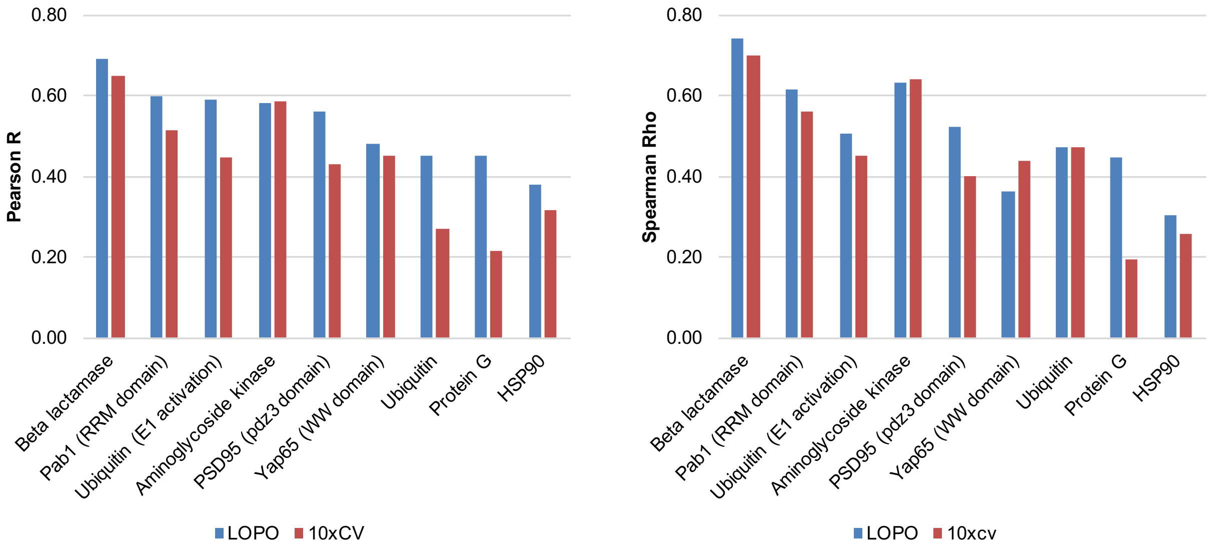
Supplementary Figure 5; related to Figure 3B. Scatter plots show the correlation between leave-one-protein-out model predictions and observed variant effects.



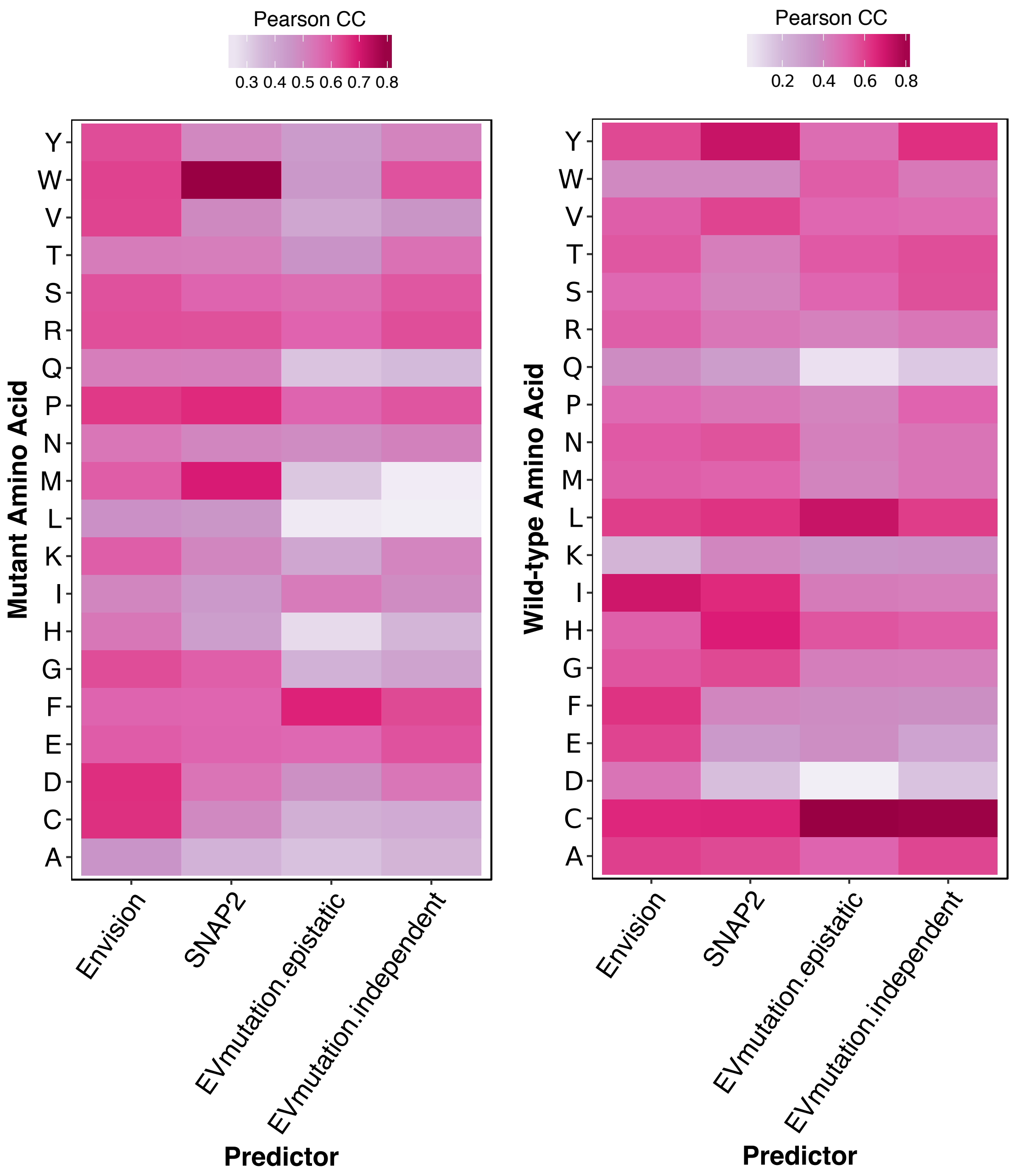
Supplementary Figure 6; related to Figure 3B. Leave-one-protein-out models were trained either with normalized or non-normalized variant effect scores. The barplots show Pearson's (left) and Spearman's (right) correlation coefficients between observed variant effect scores and predicted variant effect 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.



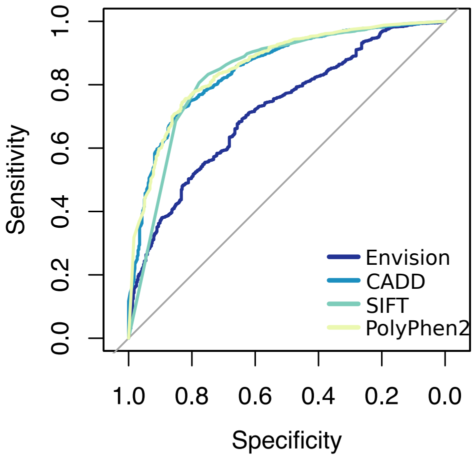
Supplementary Figure 7; related to Figure 3C. Our leave-one-protein-out models compare favorably to SNAP2 and EVmutation models. This barplot shows the correlation between predicted and observed variant effect scores for each data set for SNAP2, EVmutation (epistatic and independent models) and our leave-one-protein-out models. The x-axis shows the protein/domain withheld from training. Here, we observe that our models outperform other predictors that our models have yet to see in training.



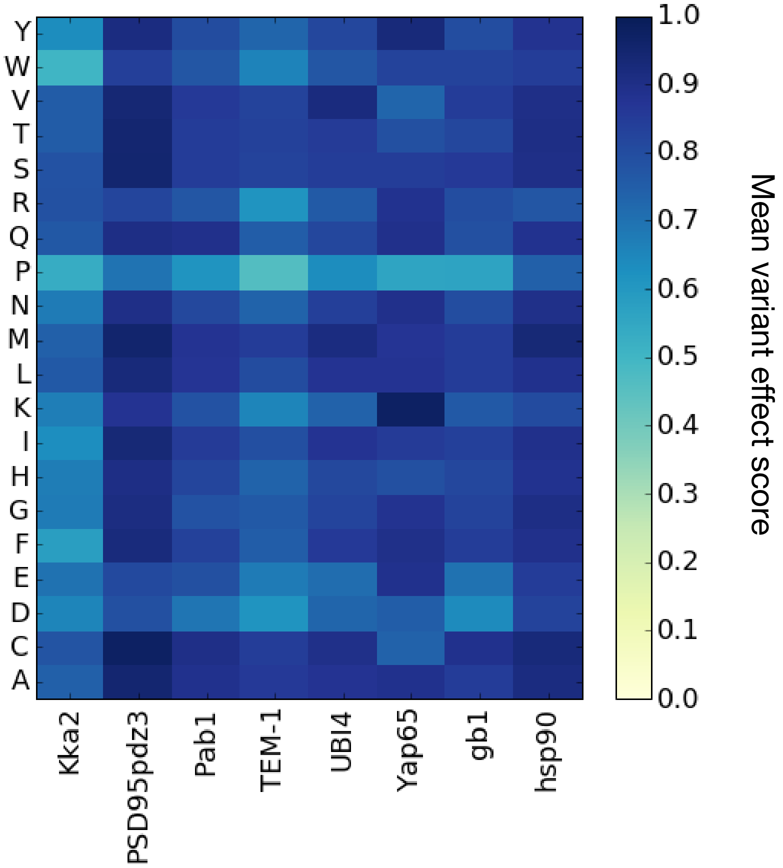
Supplementary Figure 8; related to Figure 3C. Effect of hyperparameter tuning cross-validation procedure. These barplots show the Pearson (left) and Spearman (right) correlations (y-axes) between predicted and observed variant effect scores for the left-out protein for models trained with hyperparameters 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.



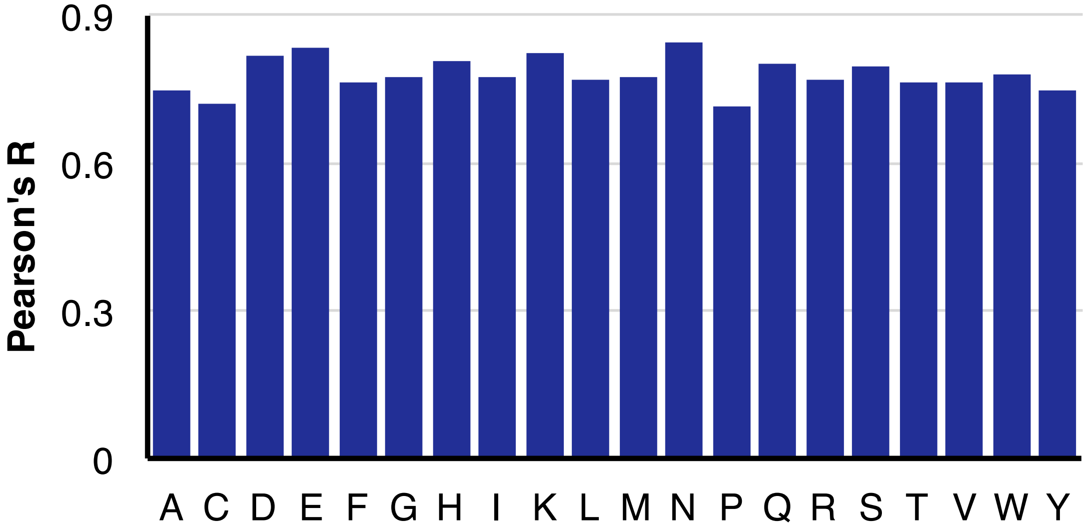
Supplementary Figure 9A-B; related to Figure 3E. Heatmaps show the correlation (Pearson's R) between predictions from four predictors for TP53 mutations arcoss mutant (**G**) and wild-type (**H**) amino acids. Darker red denotes more accurate predictions, while white shows poor predictive performance.



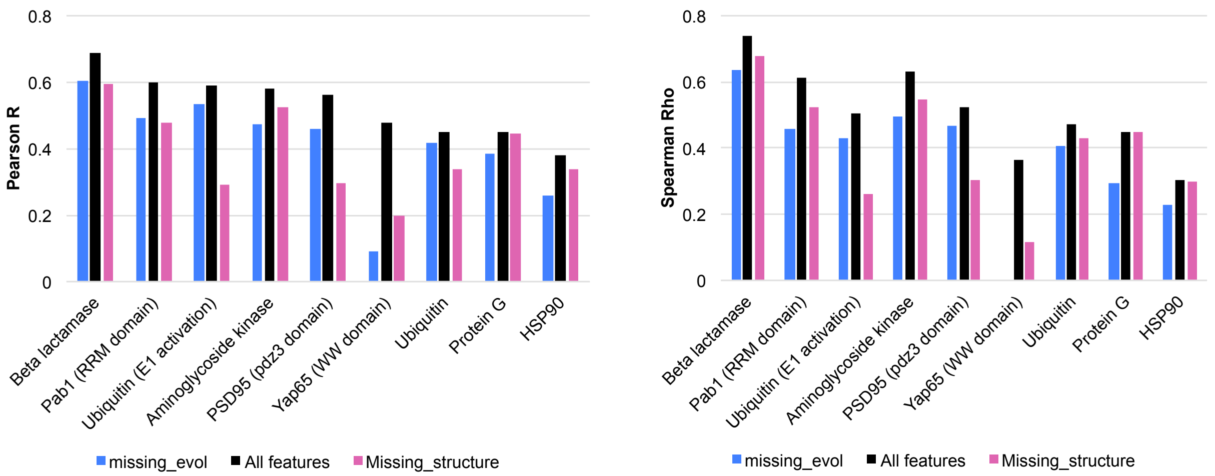
Supplementary Figure 10; related to Figure 3A. Envision, CADD, SIFT and PolyPhen2 were used to predict 9,028 pathogenic and 402 benign mutations from the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/). Receiver operator characteristic (ROC) curves were generated for each model using the pROC 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.



Supplementary Figure 11; related to Figure 4A. 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.



Supplementary Figure 12; related to Figure 4A. 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.



Supplementary Figure 13; related to Figure 2D. The leave-one-protein-out models we 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.

Supplementary Table 1; related to Figure 1A. 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 acivity (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	Q9ES00	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

Supplementary Table 2; related to Figure 1D. 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	MY 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

Supplementary Table 4; related to Figure 3A. 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	

Supplementary Table 5; related to Figure 4A. 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.