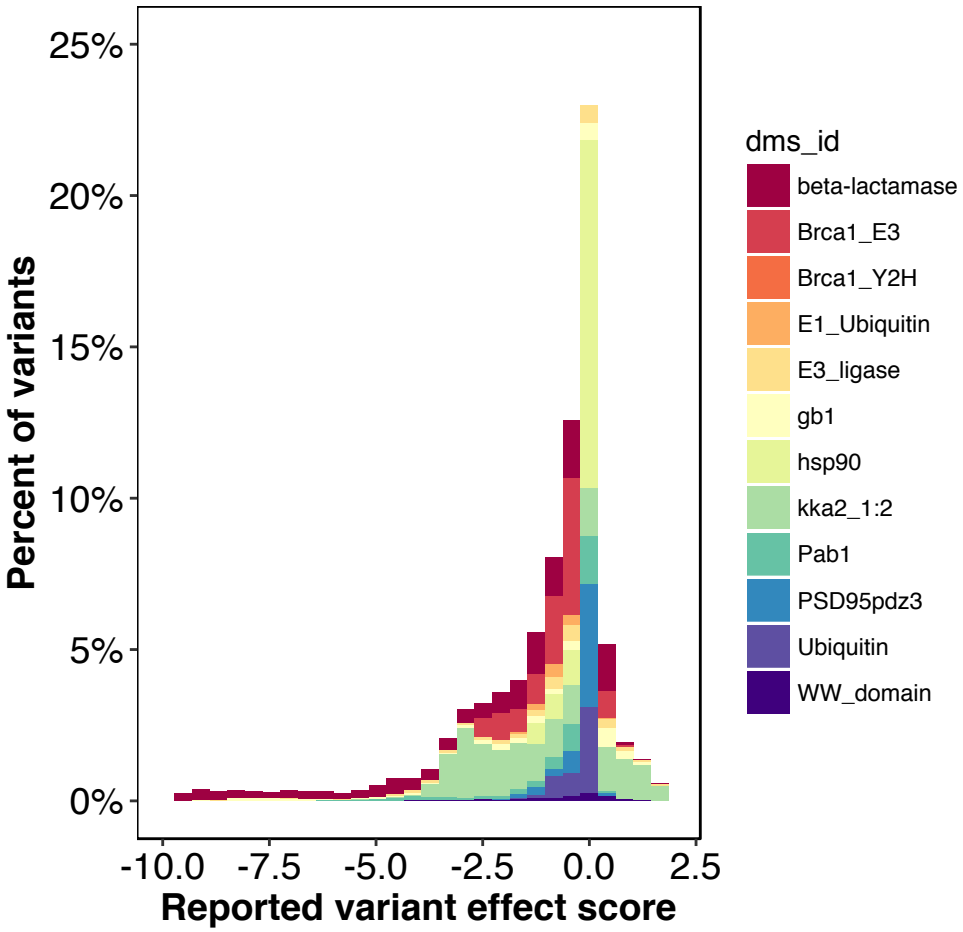
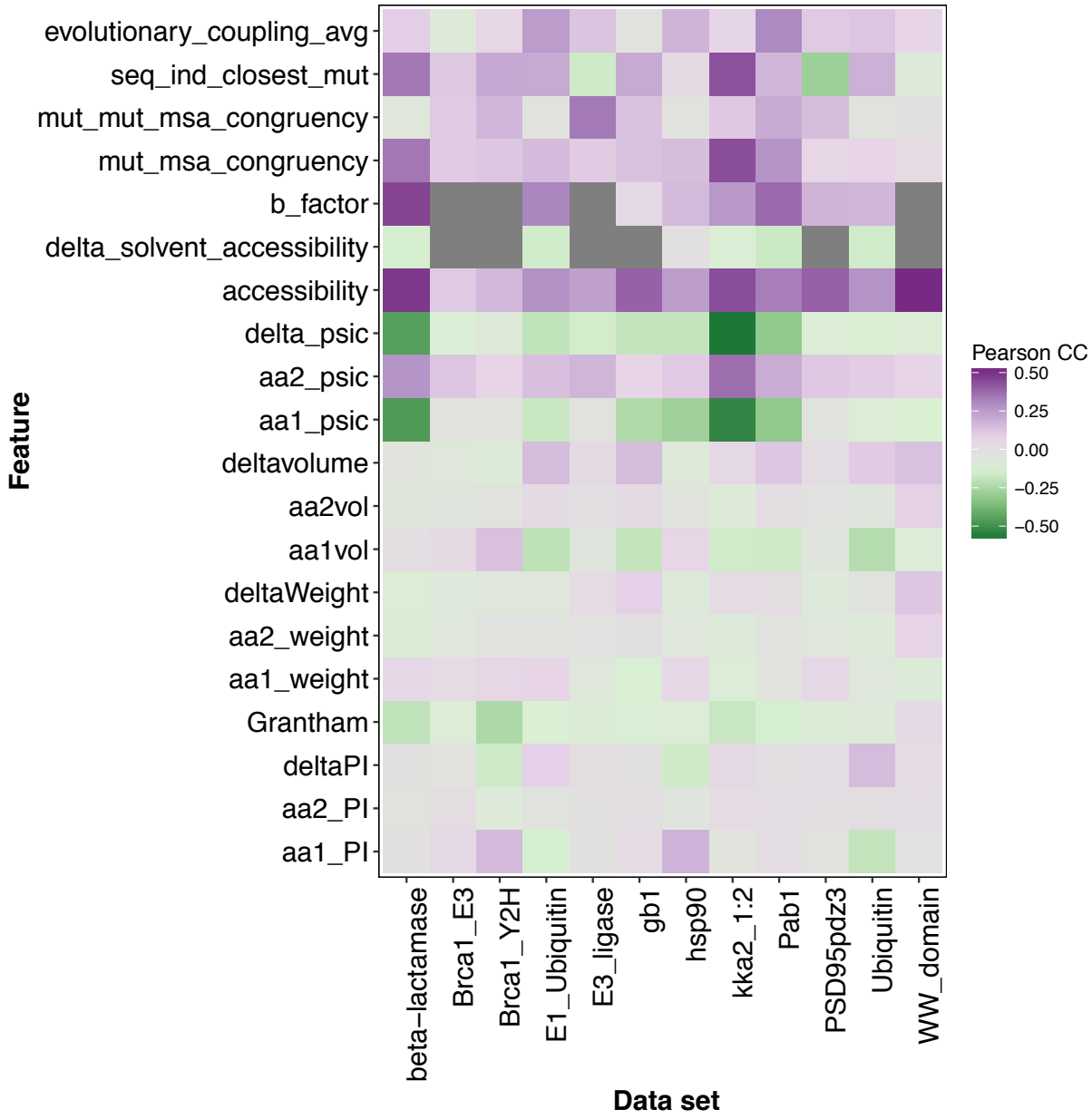
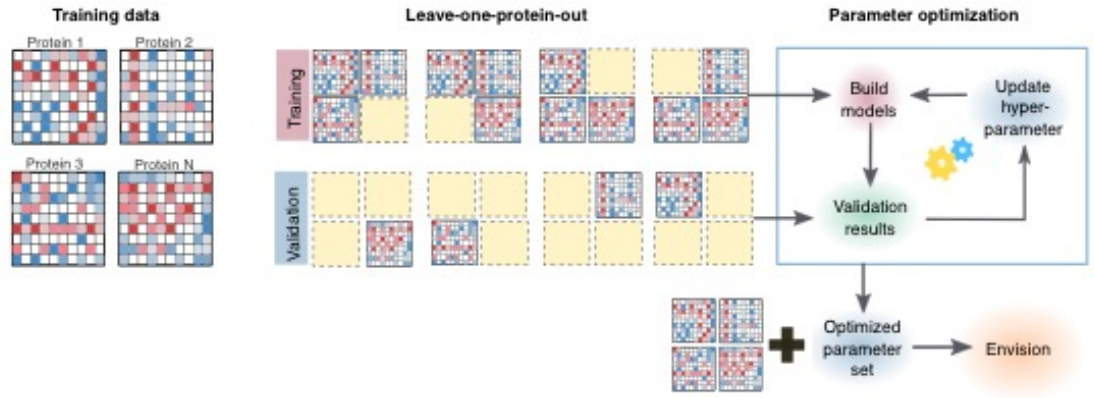


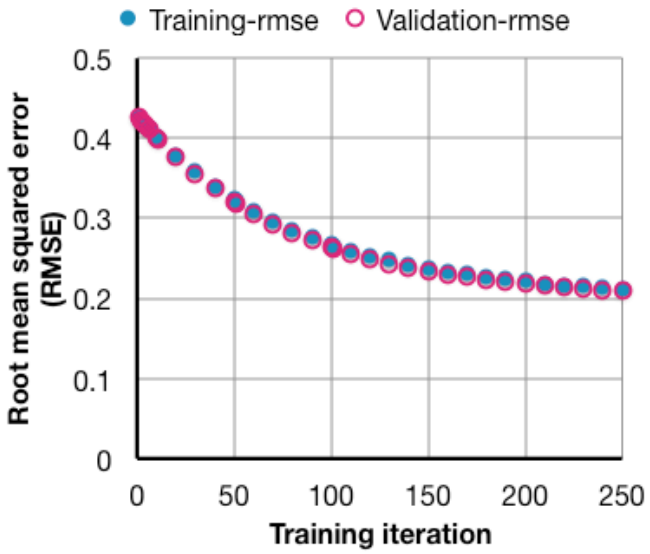

**Supplementary Figure 1; related to Figure 1C.** A histogram shows the distributions of reported variant effect scores from 12 large-scale mutagenesis data sets.



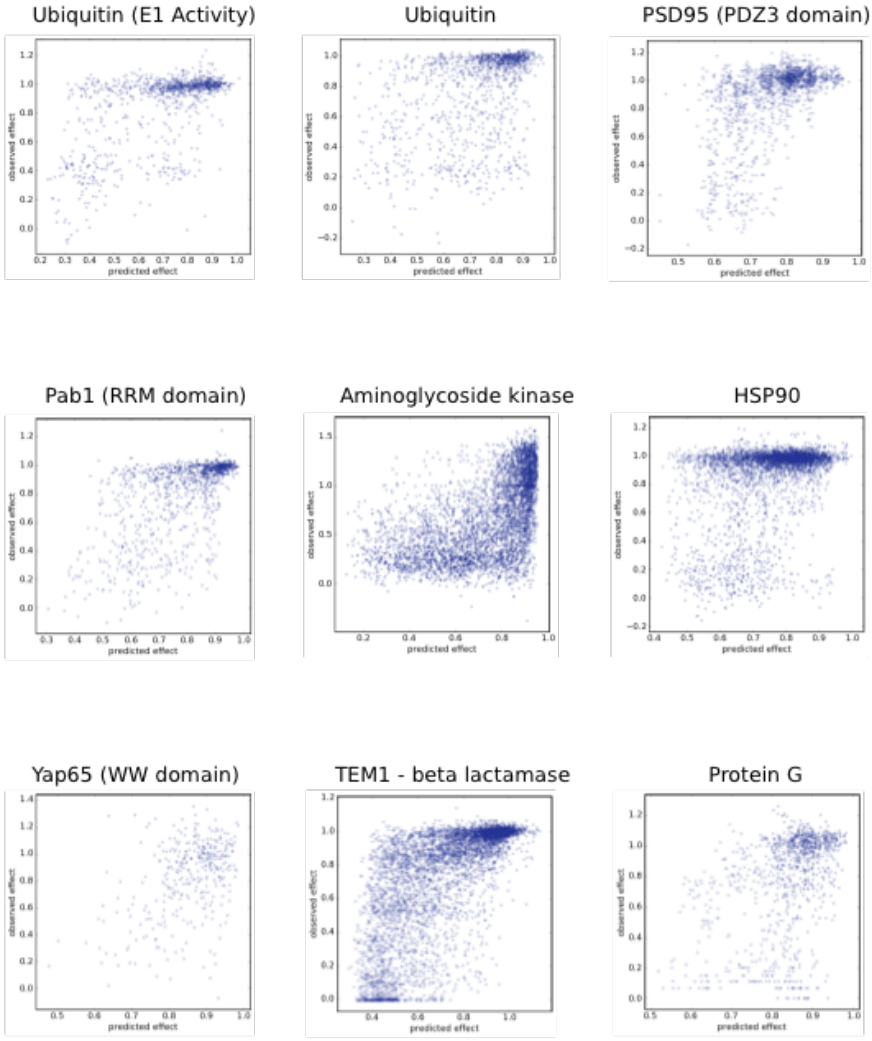

**Supplementary Figure 2; related to figure 2A.** A heatmap shows the Pearson correlation coefficient between descriptive feature values and variant effect scores for each large-scale mutagenesis data set. Note, E3 ligase, and BRCA1 datasets are missing B factor and predicted change in solvent accessibility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



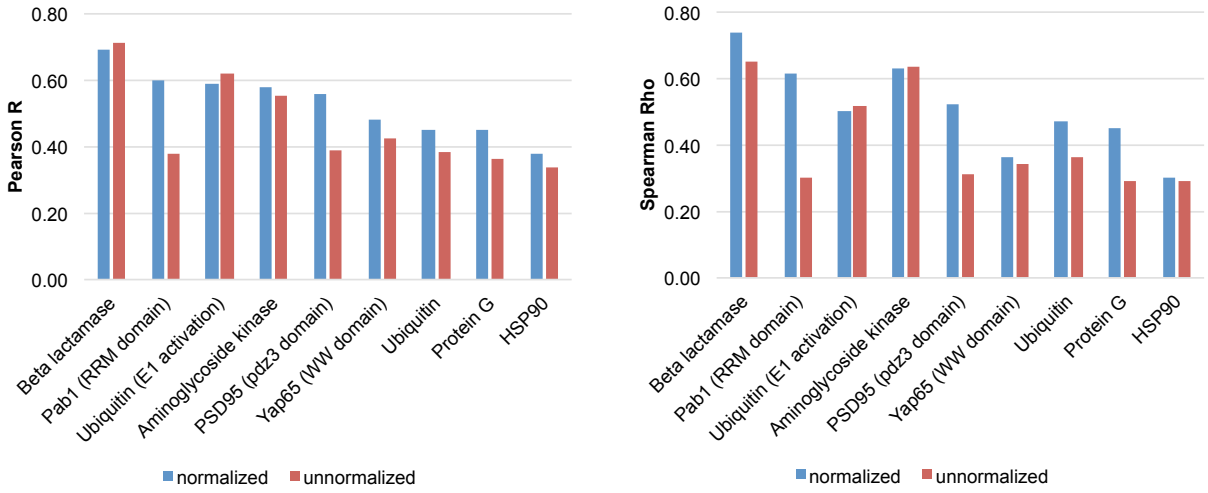

**Supplementary Figure 3; related to Figure 3A.** Our hyperparameter tuning scheme is designed to generate generalizable models. To determine the optimal values for each hyperparameter, we used a leave-one-protein-out cross-validation approach. To begin, we collected large-scale mutagenesis data sets 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.



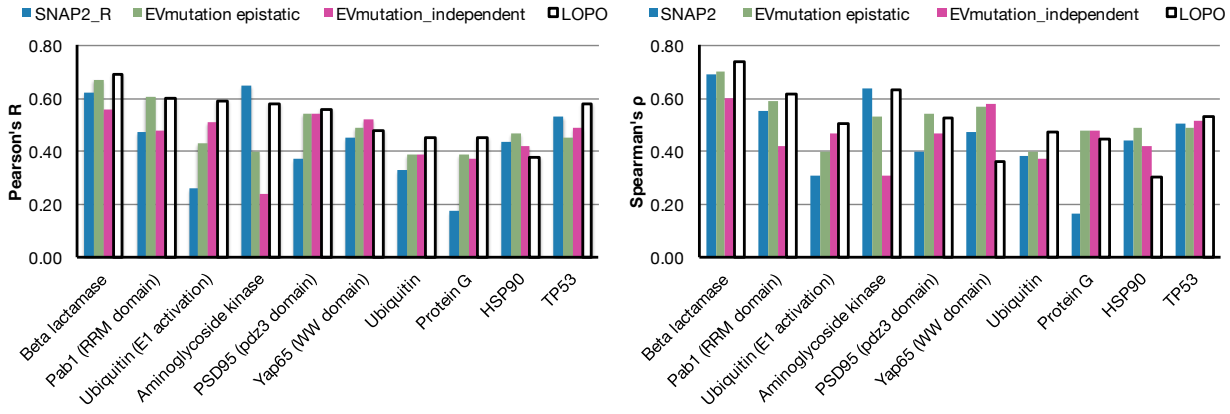

**Supplementary Figure 4; related to Figure 3A.** Training and testing data set RMSEs are very similar across iterations. While training Envision, 5% of data was withheld to determine the performance of the model as each tree was trained and added to the ensemble of decision trees. The plot shows the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



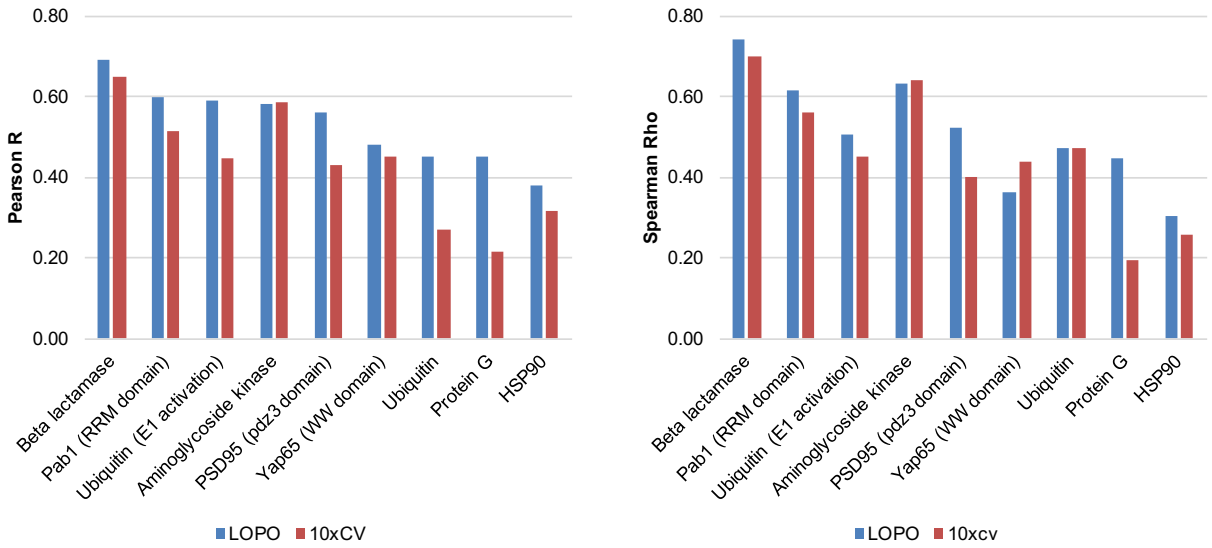

**Supplementary Figure 5; related to Figure 3B.** Scatter plots show the correlation between leave-one-protein-out model predictions and observed variant effects.



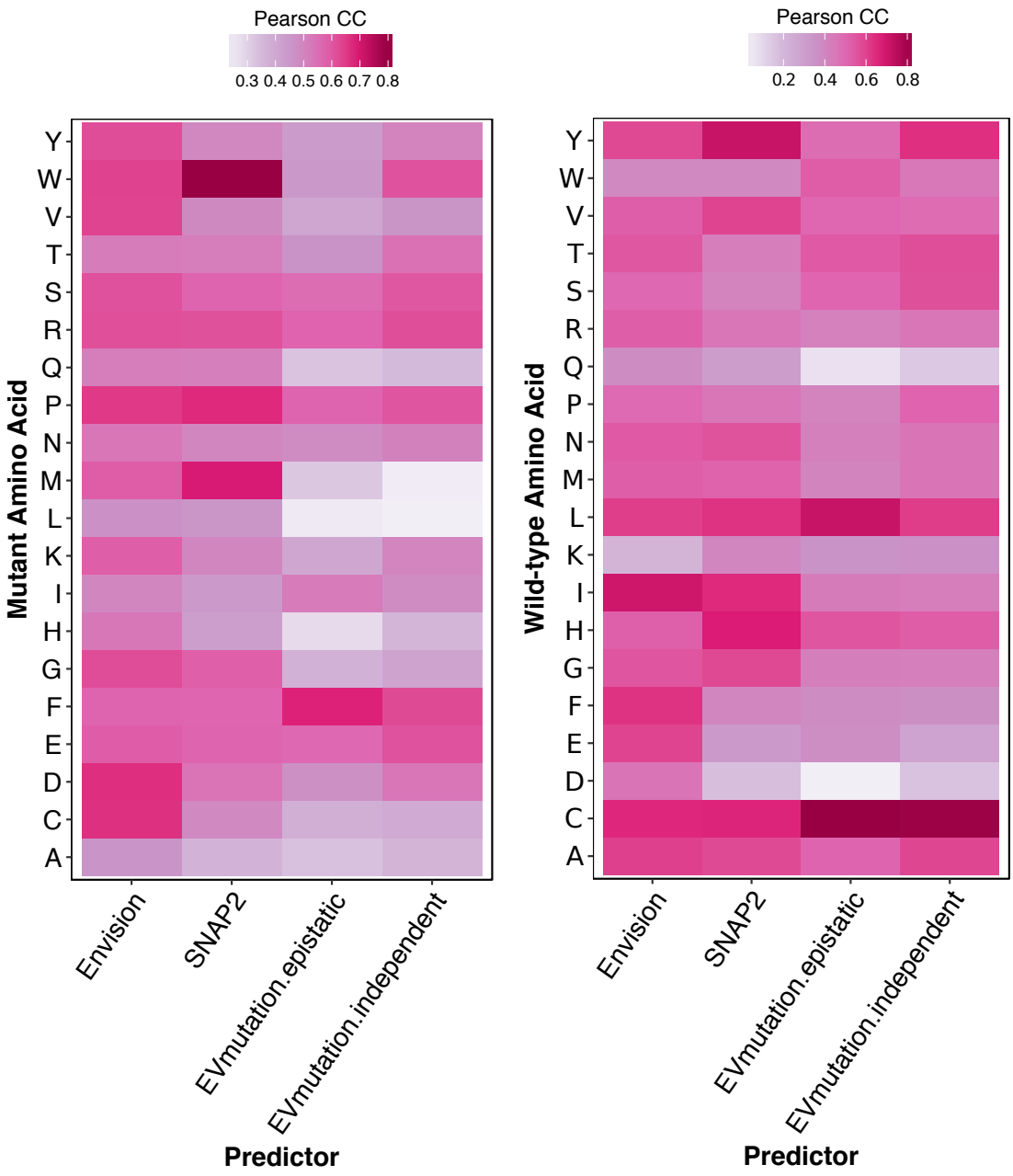

**Supplementary Figure 6; related to Figure 3B.** Leave-one-protein-out models were trained either with normalized or non-normalized variant effect scores. The barplots show Pearson's (left) and Spearman's (right) correlation coefficients between observed variant effect scores and predicted variant effect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



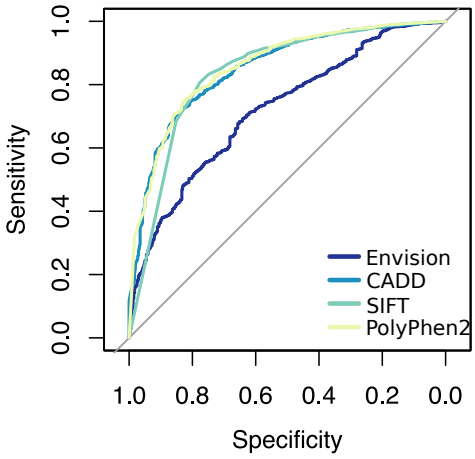

**Supplementary Figure 7; related to Figure 3C.** Our leave-one-protein-out models compare favorably to SNAP2 and EVmutation models. This barplot shows the correlation between predicted and observed variant effect scores for each data set for SNAP2, EVmutation (epistatic and independent models) and our 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.



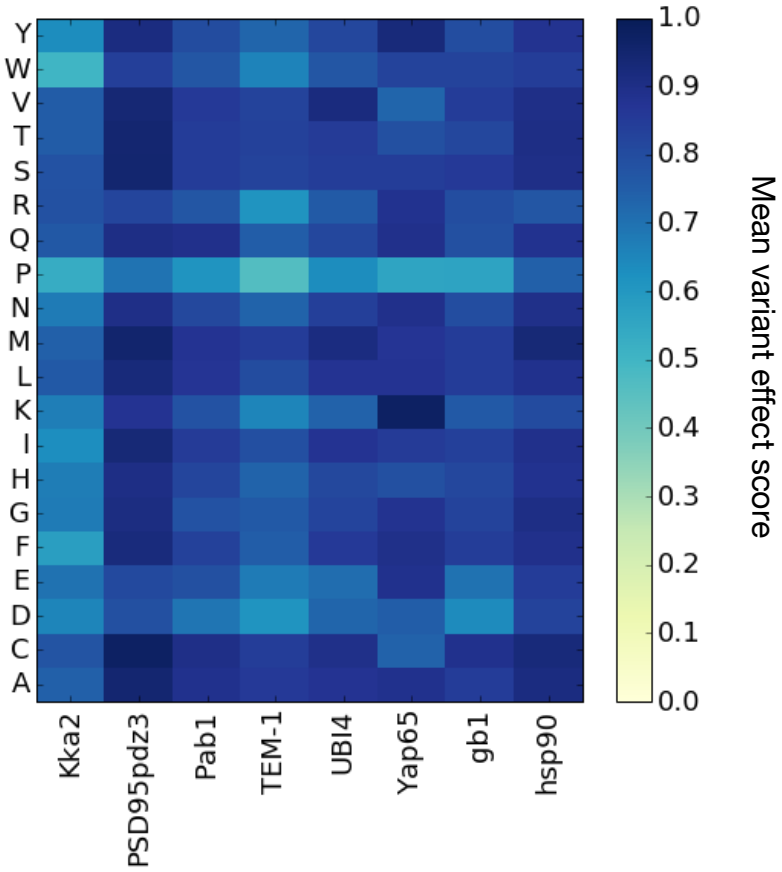

**Supplementary Figure 8; related to Figure 3C.** Effect of hyperparameter tuning cross-validation procedure. These barplots show the Pearson (left) and Spearman (right) correlations (y-axes) between predicted and observed variant effect scores for the left-out protein for models trained with hyperparameters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



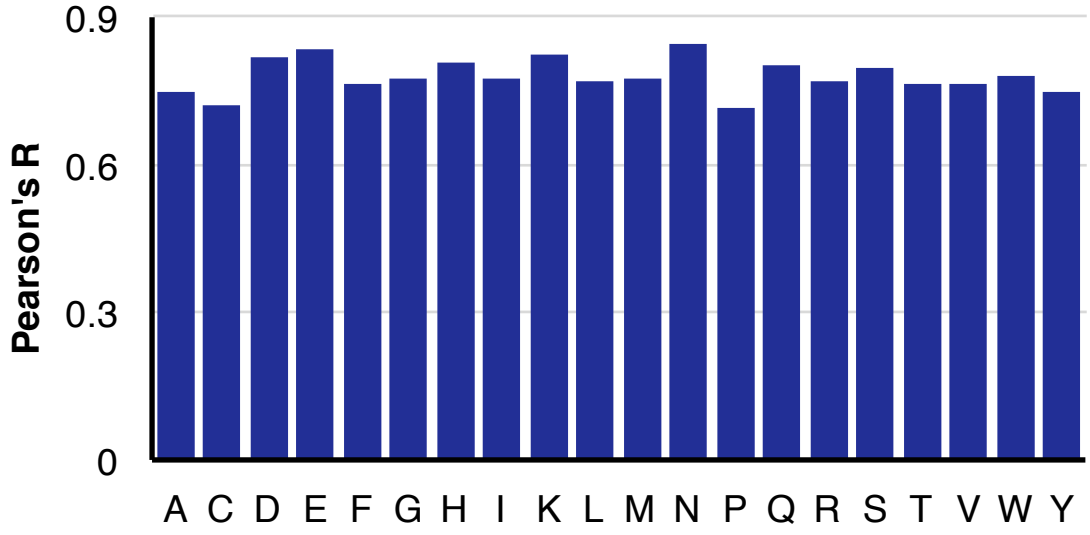

**Supplementary Figure 9A-B; related to Figure 3E.** Heatmaps show the correlation (Pearson's R) between predictions from four predictors for TP53 mutations arcoss mutant (**G**) and wild-type (**H**) amino acids. Darker red denotes more accurate predictions, while white shows poor predictive performance.



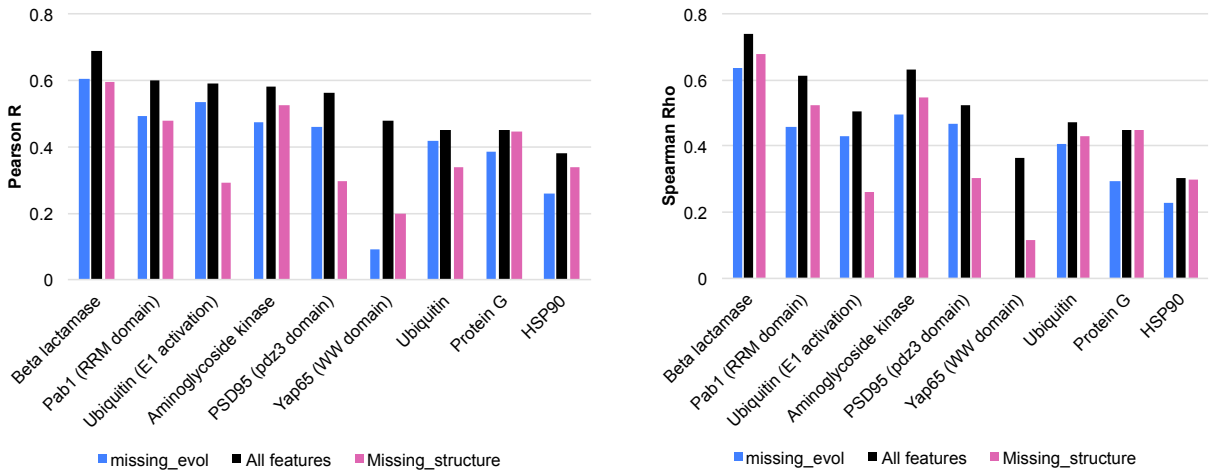

**Supplementary Figure 10; related to Figure 3A.** Envision, CADD, SIFT and PolyPhen2 were used to predict 9,028 pathogenic and 402 benign mutations from the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/). Receiver operator characteristic (ROC) curves were generated for each model using the 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 | 2E9 | 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 | Q9ESO0 | 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 | eximum 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.