Toward Developing Intuitive Rules for Protein Variant Effect Prediction Using Deep Mutational Scanning Data Supplementary Information

Cheloor Kovilakam Sruthi[1], Hemalatha Balaram[2] and Meher K. Prakash[1,*]

[1]*Theoretical Sciences Unit,* [2]*Molecular Biology and Genetics Unit,*

Jawaharlal Nehru Centre for Advanced Scientific Research,

Bangalore-560064, India

*Corresponding author : meher@jncasr.ac.in

Supplementary Figures

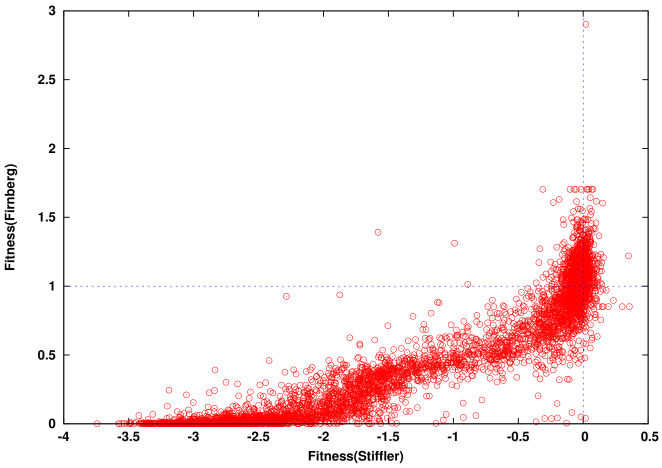

Figure S1. Comparison of the fitness effects of mutations in β-lactamase reported in two different deep mutational scan experiments. The fitness scores were obtained from Table S1 of the study by Stiffler et al.[1] and the supplementary material DataS1-S4.xlsx (Sheet S2 on missense mutations) of the study by Firnberg et al.[2] The two dashed lines passing through 0 and 1 represent the wild-type fitness in the Stiffler et al.[1] and Firnberg et al.[2] experiments, respectively.

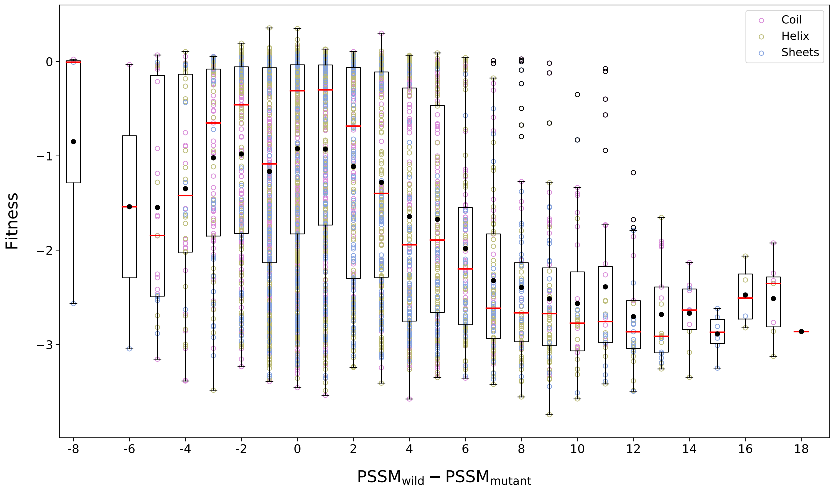

Figure S2. Comparison of fitness with $\Delta PSSM$ ($= PSSM_{wildtype} - PSSM_{mutant}$). Amino acid conservation at a site captures only the average fitness effect upon substitution and does not give information about specific substitutions. To capture the fitness effect of each substitution at a site, we use the Position-Specific Scoring Matrix (PSSM) scores which are based on the amino acid frequencies observed in the multiple sequence alignment (MSA) of the protein. The PSSM was calculated using PSI-BLAST for the MSA of β-lactamase. The fitness effects show a dependence, albeit weak.

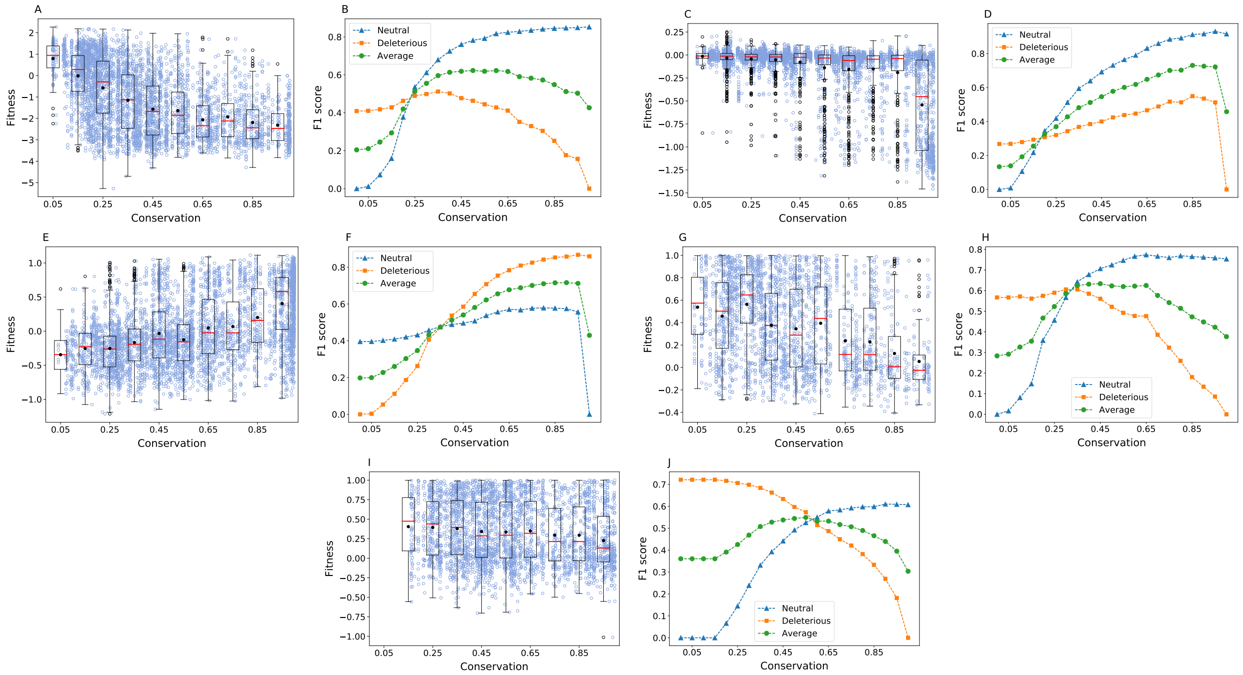

Figure S3. Fitness dependence on the amino acid conservation at a position obtained from the multiple sequence alignment (MSA) of the protein. Using a threshold for conservation the substitutions are classified as neutral and deleterious, and the quality of classification as quantified by F1 score at different values of conservation threshold is shown for the proteins APH(3′)-II **(A,B)**, Hsp90 **(C,D)**, MAPK1 **(E,F)**, UBE2I **(G,H)** and TPK1 **(I,J)**. In the box plots, the black filled circle and the red line represent the mean and median of the fitness, respectively. The whiskers are plotted at the lowest data point greater than Q1-1.5×(Q3-Q1) and the greatest data point less than Q3+1.5×(Q3-Q1) where Q1 and Q3 represent the first and the third quartile, respectively. Black open circles show the outliers.

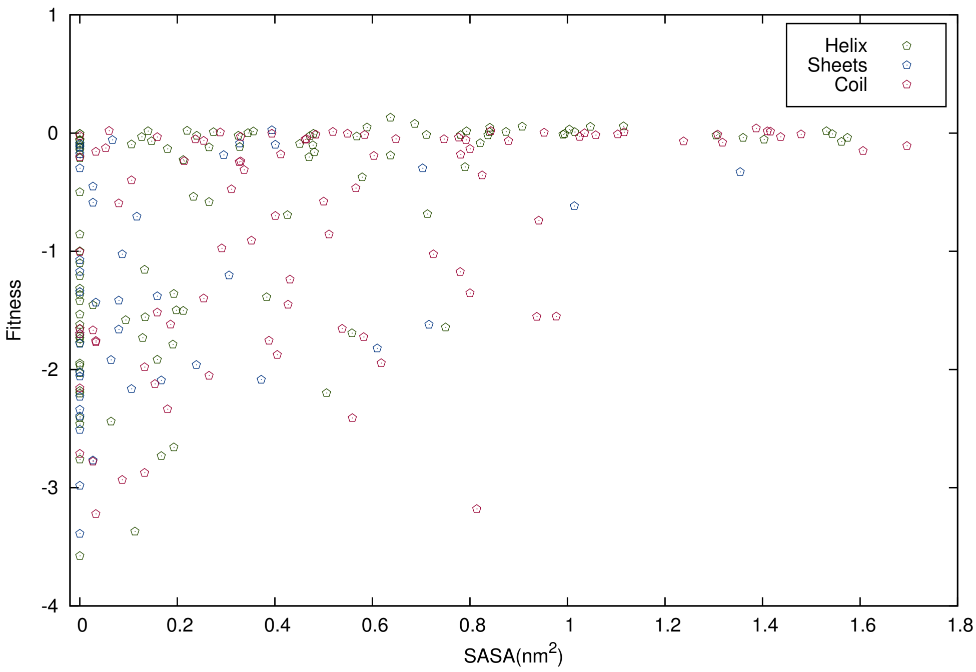

Figure S4. For the sake of simpler interpretations the quantitative relation of fitness with SASA only for alanine substitutions in β-lactamase is examined. The analysis reflects triangular pattern with solvent exposure.

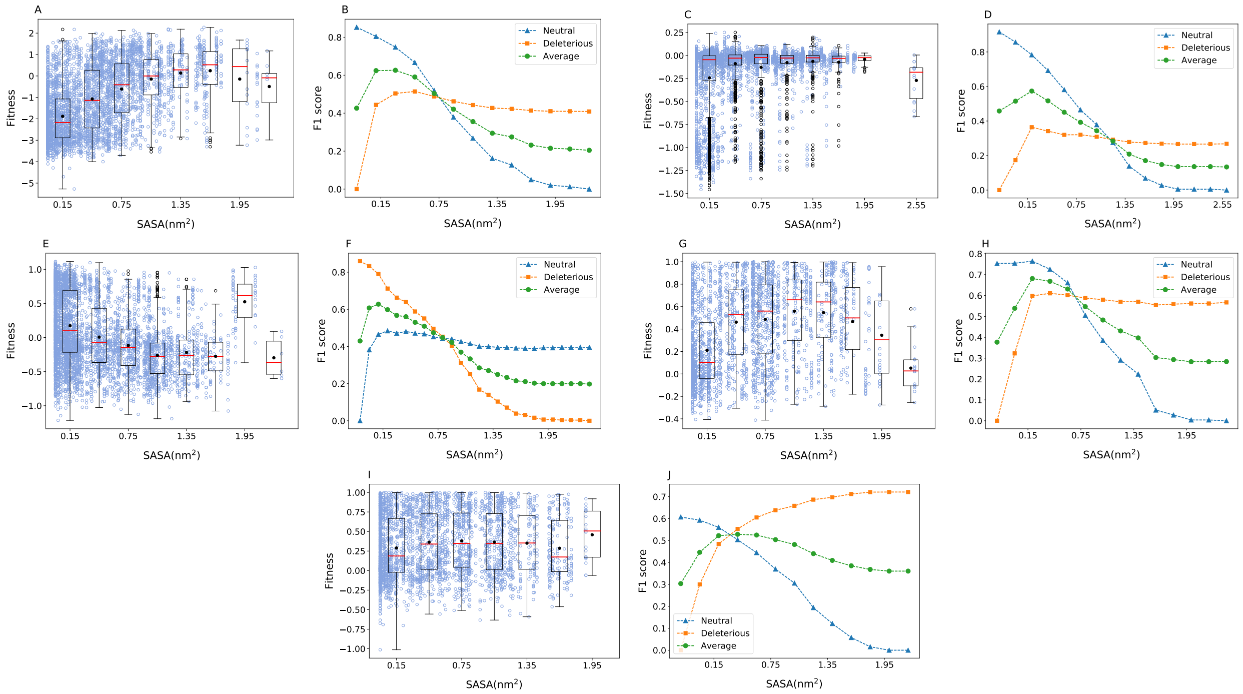

Figure S5. Correlation of fitness with SASA and the F1 scores quantifying the quality of classification as the SASA threshold is varied for the proteins APH(3′)-II **(A,B)**, Hsp90 **(C,D)**, MAPK1 **(E,F)**, UBE2I **(G,H)** and TPK1 **(I,J)**. Details of the box plot representation are given in **Figure S3**.

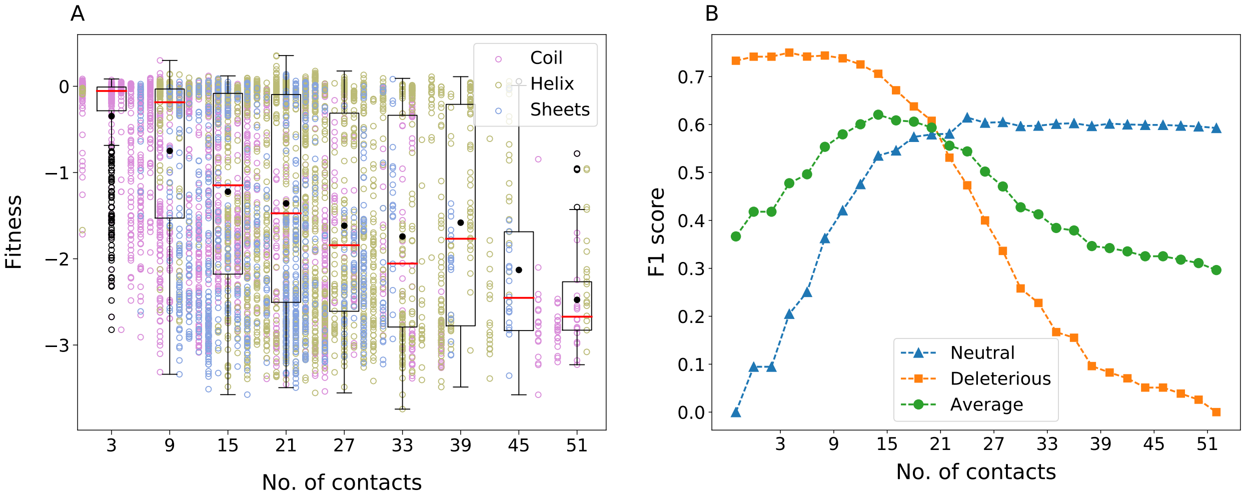

Figure S6. **A.** The effect of packing was studied by plotting fitness relative to the number of contacts. No strong relation was observed. **B.** F1 scores for neutral and deleterous classes and the average of both as the number of contacts threshold is changed. For details of box plot representation see **Figure S3**.

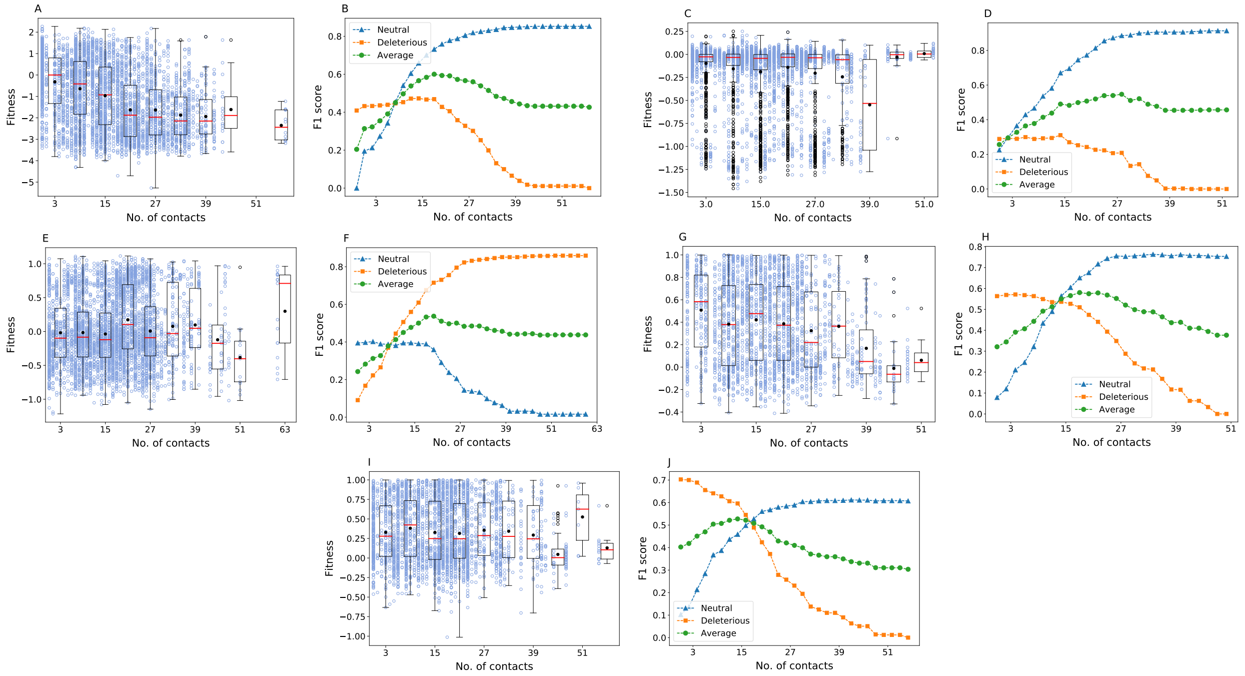

Figure S7. Fitness scores with respect to the number of contacts of the wild-type amino acid and the F1 scores at different number of contacts thresholds chosen for classification. Data is shown for the proteins APH(3′)-II **(A,B)**, Hsp90 **(C,D)**, MAPK1 **(E,F)**, UBE2I **(G,H)** and TPK1 **(I,J)**. (Box plot representation details are given in **Figure S3**.)

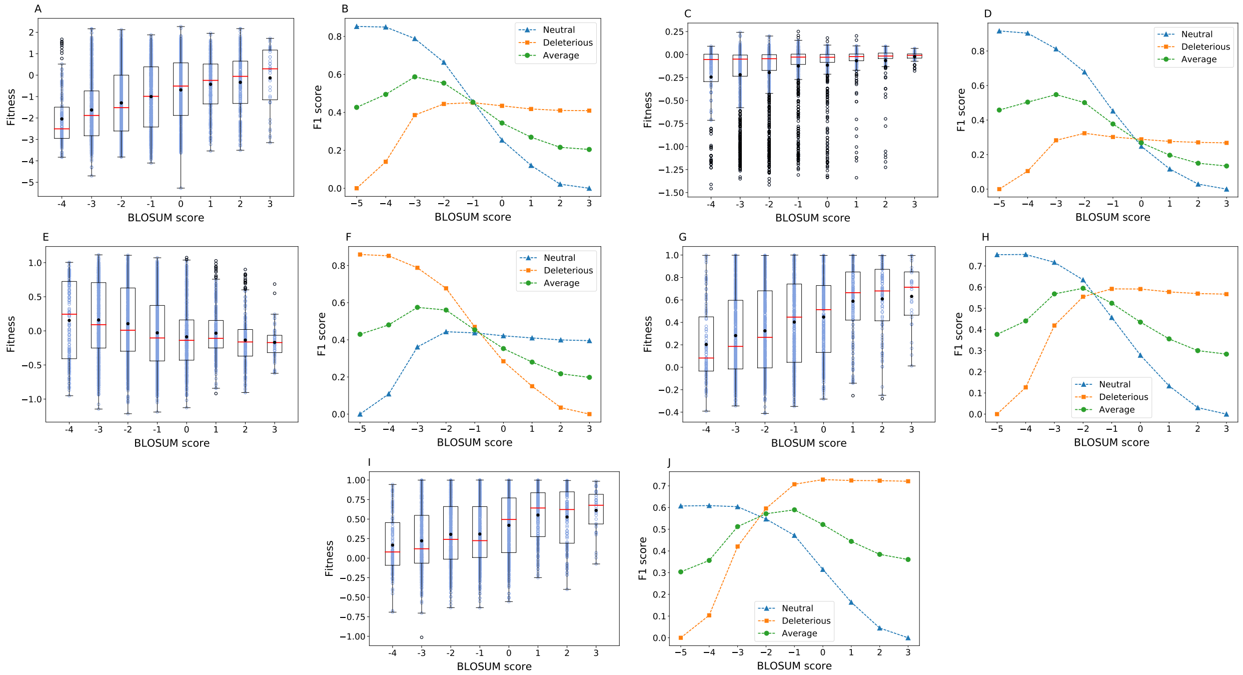

Figure S8. Correlation between fitness and BLOSUM substitution matrix score and the F1 scores when the BLOSUM threshold for classifying mutations to neutral and deleterious is varied for the proteins APH(3′)-II **(A,B)**, Hsp90 **(C,D)**, MAPK1 **(E,F)**, UBE2I **(G,H)** and TPK1 **(I,J)**. See **Figure S3** for details of box plot representation.

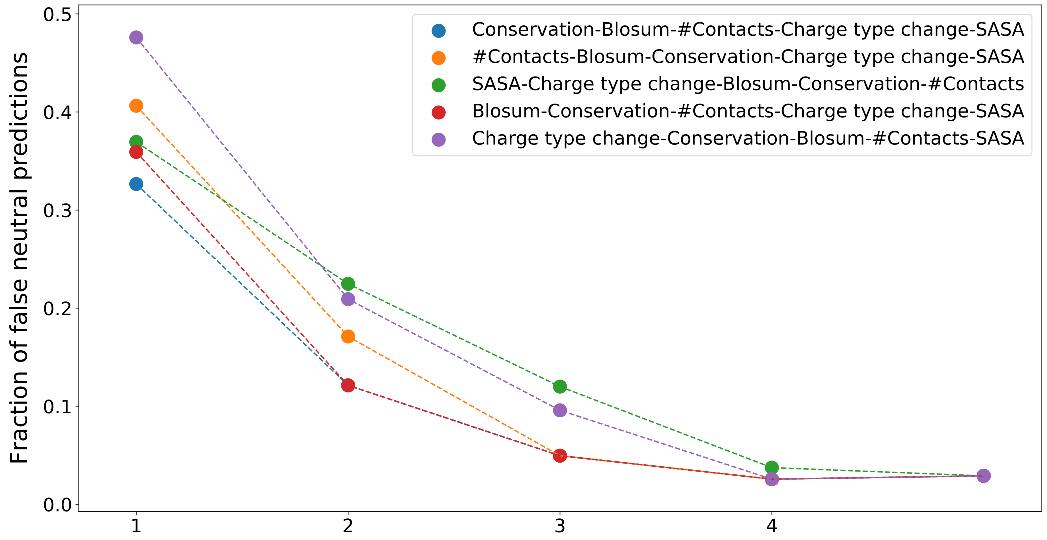

Figure S9. Reduction in the number of mutations predicted wrongly as neutral as the number of criteria used is increased is shown for the case of β-lactamase. The order in which thresholds related to different variables are included is shown in the legend.

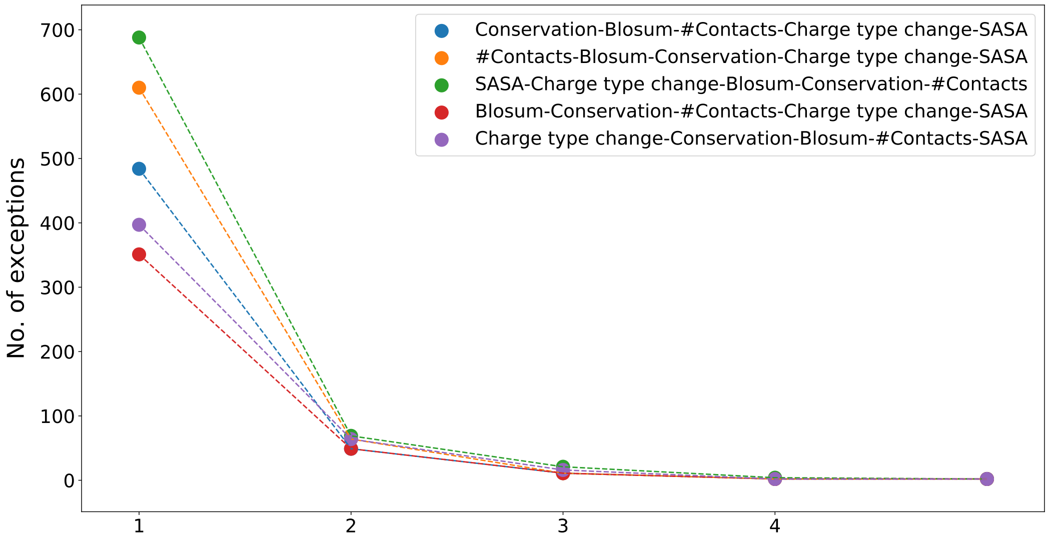

Figure S10. Reduction in the number of mutations predicted incorrectly as neutral as the number of variables used is increased. Although the error fraction decreases as shown in **Figure S9**, the number of mutations which are classified as neutral or deleterious by all variables also decreases.

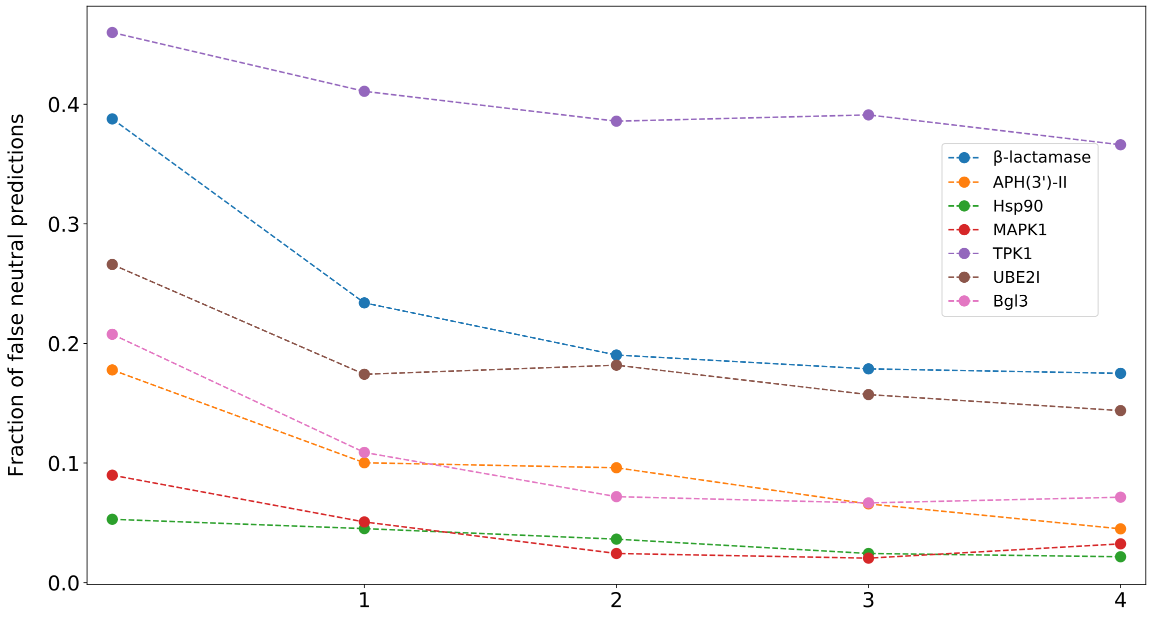

Figure S11. Reduction in the chance of false-neutral predictions as the number of variables used for classification is increased. Threshold used for each variable is the average values given in **Table 1**. The trajectory for which the sum of error fractions is the least is shown for each protein. The order in which different threshold criteria are included for each protein is: 1) β-lactamase: SASA-Charge type change-Blosum-No.of contacts-Conservation, 2) APH(3′)-II: Blosum-No. of contacts-Charge type change-SASA-Conservation, 3) Hsp90: Conservation-Charge type change-SASA-Blosum-No. of contacts, 4) MAPK1: Conservation-Charge type change-SASA-Blosum-No. of contacts, 5) TPK1: Charge type change-SASA-Conservation-No. of contacts-Blosum, 6) UBE2I: SASA-Charge type change-No. of Contacts-Blosum-Conservation, 7) Bgl3: SASA-Charge type change-No. of contacts-Blosum-Conservation

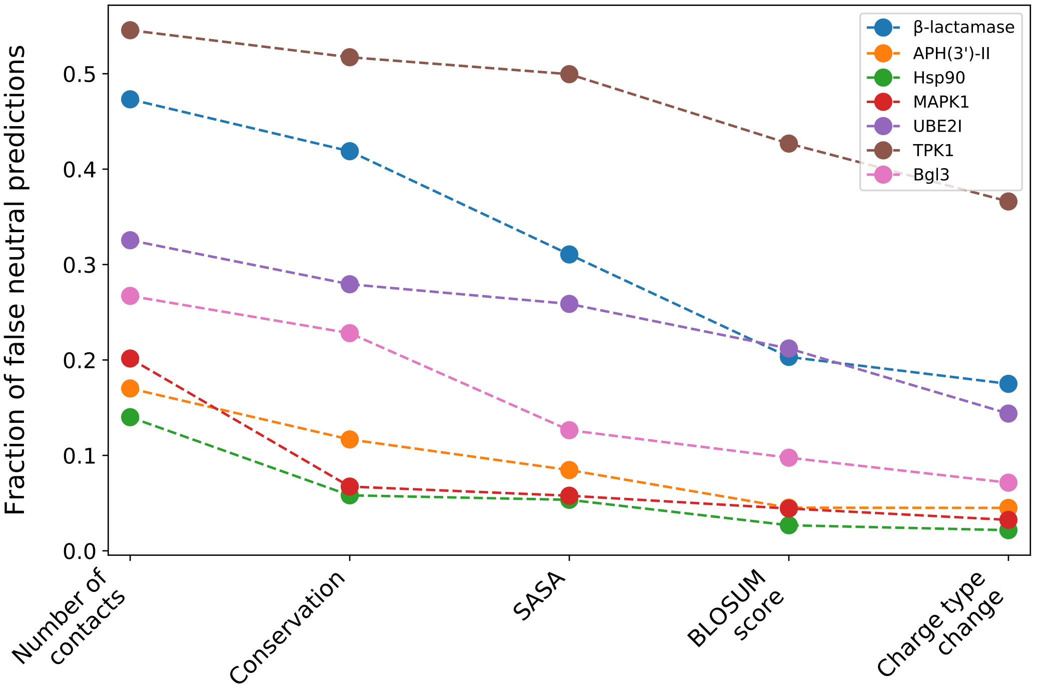

Figure S12. Variation in the chance of false-neutral predictions on icreasing the number of threholding criteria used for classification. Here the drop in error is shown for the specific order of variables: Number of contacts-Conservation-SASA-BLOSUM-Charge type change. Threshold used for each variable is the average values given in **Table 1**.

[1] Stiffler, M. A.; Hekstra, D. R.; Ranganathan, R. Evolvability as a function of purifying selection in TEM-1 β-lactamase. *Cell* **2015**, *160*, 882–892.

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