Interpreting mutational effects predictions, one substitution at a time

C. K. Sruthi and Meher K. Prakash*
Theoretical Sciences Unit
Jawaharlal Nehru Center for Advanced Scientific Research,
Jakkur,
Bangalore 560064
India

Corresponding author: meher@jncasr.ac.in

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

Keywords: Explainable AI, Interpretable AI, mutational effects, deep mutational scan

Introduction.

Proteins perform the most critical cellular functions, and yet they are very delicate in their design. A small perturbation to the sequence such as a change in a single amino acid of the protein, known as a mutation, can alter their structure and function significantly. In fact, several of the disease phenotypes such as cancers, Alzheimer's [1] as well as other problems associated with infectious diseases such as antibiotic resistance [2] can all be traced back to mutations in proteins. Understanding the effects of mutations is also relevant in protein engineering, where one would like to introduce the changes that affect the solubility and function of the proteins in a predictable way.

One of the ways of understanding mutational effects is by introducing these mutations in the proteins [3], in a site-directed or random fashion, and studying the structural or functional consequences of it. In a commonly practiced protocol, known as the alanine scan [4], amino acids of interest are substituted by alanine, which is small and neutral, and the functional consequences are studied. In recent years, technological advances allowed a much more detailed scan, known as the deep mutational scan (DMS) [5-7]. In DMS, each amino acid is replaced by all possible 19

amino acids, and cellular level consequences are studied. The exhaustive sampling allows thousands of mutations to be studied at a single substitution level, and hundreds of thousands with two to three substitutions.

The computational efforts to understand the mutations are also making significant strides. Artificial Intelligence based models are useful for complementing the missing data in DMS [8], or making detailed predictions of the complete mutational scan using co-evolutionary relations [9] or by training on a few DMS experiments [10]. The models trained on large mutational databases curated from across the literature [11] or from a few DMS studies [10] are making good predictions of the mutational effects.

Ability to make reliable mutational effects predictions brings one to a natural point of asking the questions that are being asked in the machine learning community, such as why the predictions should be trusted [12] or alternatively if the predictions can be interpreted [13,14]. The philosophical debate about predictability versus interpretability [15] is being revisited in several areas of machine learning, specifically when one is interested in controlling the effects by developing an understanding for the contributing factors.

We introduce this notion of interpretability or explainability in the context of mutational effects prediction. Interpretability begins with a small shift in perspective, from asking how do various factors contribute to the *set of predictions*, to what is the contribution of the various factors to an *individual prediction*. In a linear regression model, knowing the measured outcome, it is trivial to understand the relative contributions from the different factors. The same is not true while working with machine learning models, which have non-trivial and non-explicit relations between the inputs and the outcome.

We use a method called SHAP (SHapley Additive exPlanations) [16,17] that is being used in several areas of machine learning, to interpret the contributions to the mutational effects. SHAP is based on the game theoretical questions raised by Shapley [18] about how the gain can be shared by different contributing players. In SHAP, the feature contributions are additive, thus making their relation to the outcome easy to interpret. We apply SHAP to interpret the outcomes of fitness [19] and solubility [20] in the deep mutational scans of β-lactamase protein. The interpretability allows us to revisit the classical intuitions on how different factors can influence mutations with a quantitative perspective.

2. Methods.

2.1 Deep mutational scan data. The deep mutational scan data of the mutational effects of β-lactamase on fitness was obtained from Stiffler et al. 2015 [19] and solubility from Klesmith et al. 2016 [20]. We use the yeast surface display (YSD) data on the effects of mutations on solubility [20] and the changes in relative fitness of *E. Coli* when a mutant containing strain is challenged with 2500 μg/ml ampicillin [19]. The analyses on the solubility were presented for the substitutions on positions 61-

215 [20], with protein data bank identity 1M40 [21], and for consistency, we chose to work with the same set of mutations both for the solubility and cellular fitness.

2.2. Descriptive parameters for AI model. We used 18 parameters to describe the amino acid before substitution or the nature of the substitution [22]:

Structural factors: *1.* SASA - Solvent accessible surface area, *2.* SS - Lack or presence of order in secondary structure (0 or 1) *3.* Contacts - number of neighboring amino acid atoms within 4 Å of the wild type amino acid *4.* Catalytic_dist, which is the distance of the amino acid from the catalytic site *5.* Av_commutetime, which is the average commute time reflecting the connectivity of the amino acid to the rest of the protein [23]

Sequence factors: *6.* BLOSUM substitution matrix [24], *7.* Wt_hb, hydrophobicity of the wild type amino acid according to Kyte-Dolittle scale, *8.* Mut_hb, hydrophobicity of the amino acid after mutation *9.* PSSM_w representing the position specific substitution matrix of the wildtype amino acid and *10.* PSSM_m of the mutant, *11*. Conservation – of the amino acid across the multiple sequence alignment

Co-evolutionary factors: *12.* Av_corr, average co-evolutionary correlation of an amino acid with all other amino acids, *13.* Degree, *14.* Betweenness, *15*. Closeness, and *16.* Eigenvector centralities which are the different centrality measures defined on the symmetric co-evolution network, *17*. Impact and *I8*. Dependence defined as asymmetric co-evolutionary parameters [25]

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

Interpretable AI model. SHapley Additive explanation (SHAP) uses the formalism where an explanation model g is defined in terms of the parameter set z'_i defining each instance (in our case each individual mutation) and their corresponding additive contribution weights ϕ_i [16,17]

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

The explanatory model is subject to three conditions known as:

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

missingness – which ensures that if a variable is z'_i=0, then the weight corresponding to it is ϕ_i=0

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

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

3 Results and Discussion.

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

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

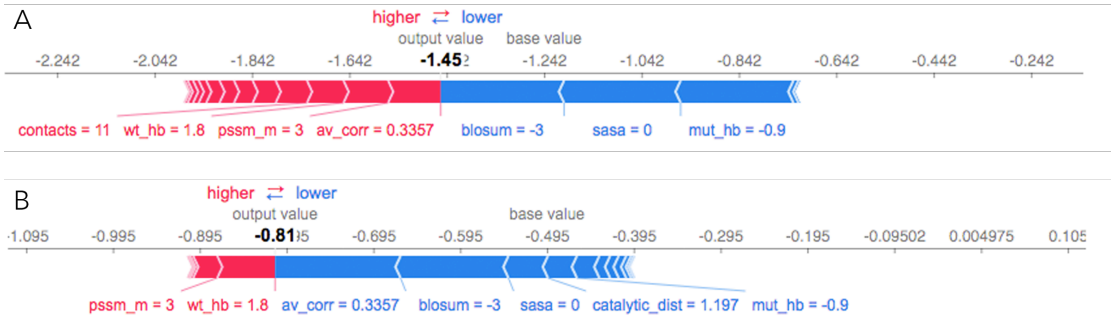


Figure 1. Decomposing the contributions. Illustration of the contributions of various factors to the effects of mutating Alanine in position 79 to Tyrosine (A79W) in β-lactamase: **A.** Fitness effect **B.** Solubility effect. As indicated by the direction of the arrows, the factors in pink contribute to an increase in the fitness or solubility and those in blue have the opposite effect. Whether a specific factor tends to increase or decrease the mutational effect depends on the individual case. The descriptive parameters are labeled along with the values they assume in this specific instance, for the specific mutation. The illustrations are generated using the Python implementation of SHAP ((https://github.com/slundberg/shap)).

3.2 Summarizing the contribution of the individual factors in the complete set.

The impact obtained from individual factors is then summarized to understand the variables that have the most significant role in the set of predictions (Figure 2).

Observing the range of the values assumed, one can infer that conservation, SASA, BLOSUM, along with the hydrophobicities of the wild type and mutant amino acids contribute significantly to both solubility and fitness. However, one can notice in the illustrations that at times the distribution of the data points labeled pink (higher outcome) and blue (lower outcome) for the some of the variables is different when comparing fitness and solubility (Figures 2A and 2B).

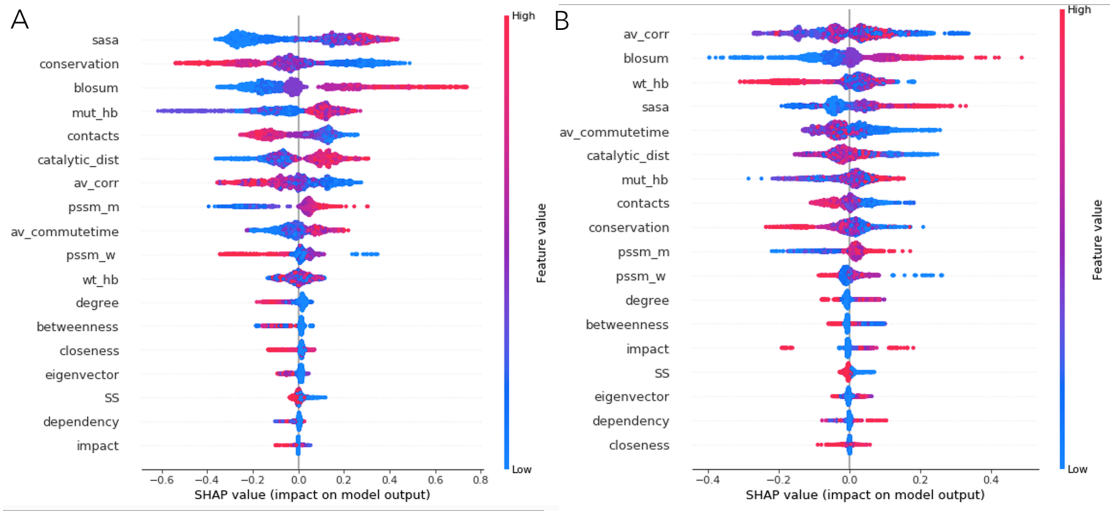


Figure 2. Summarizing the contributions. We analyzed the deep mutational scan data where the consequences of any of the 19 possible amino acid substitutions at each of the positions (61-215) were measured. The individual contributions to **A.** fitness and **B.** solubility obtained from each of these mutations are summarized in the plot. Along each line, one finds the name of the descriptive parameter, a distribution of the SHAP values across the complete set of mutations, along with the color indicator of the fitness/solubility outcome associated with mutation.

3.3 Identifying factors making correlated contributions to fitness and solubility.

We further examined the contributions of each of the parameters to fitness and solubility in the same analysis, to see if they are correlated, anti correlated or uncorrelated. As shown in Figure 3, the contributions from three variables, conservation, BLOSUM and number of contacts have a good positive correlation. The contributions of other variables to the fitness and solubility do not have strong correlations (Supplementary Figures 2A, 2B).

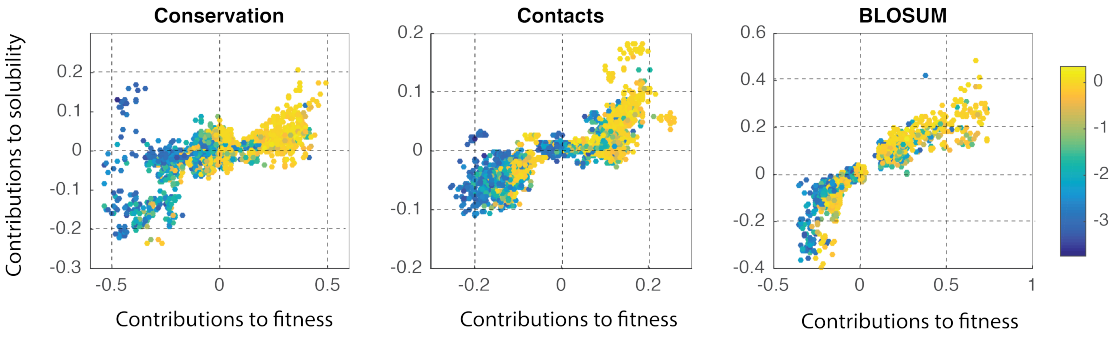


Figure 3. Correlation of contributions to solubility and fitness. The SHAP values defining the contribution of each variable to fitness and solubility are shown. From the data it is clear that the contributions from these three

descriptive parameters, conservation, number of contacts and BLOSUM are mostly correlated i.e., if the parameter contributes to an increase in fitness it also contributes to an increase in the solubility. This is not true for all variables, which are shown in the Supplementary Figures 2 and 3, where they are uncorrelated or negatively correlated. The color bar represents the experimentally observed fitness changes among all the mutations studied.

3.4 Extracting intuitive patterns about the effect of different factors. There are several intuitions about how factors such as conservation or distance from the catalytic site should influence the protein function or its solubility. Many of these intuitions still need to be quantified. Klesmith et al. 2016 [20] performed a very careful analysis using naïve Bayesian classification to clarify the chance that a mutation characterized by a parameter is statistically likely to be deleterious or neutral. They observe an interesting trade-off such as that the distance from the catalytic site has opposite effects on solubility and fitness. We ask if one can go beyond the classification models to quantify these dependencies using SHAP analysis. The scatter plots in Figures 4A, 4B show that the relation of fitness and solubility to conservation is not easy to infer. Naively, while observing the relation of the outcomes to a single variable in a multi-factorial system, one may expect either a poor correlation or even a lack of it. However, when we plot the contributions to fitness and solubility using the SHAP analysis (Figures 4C, 4D), the dependence of the component becomes much more predictable. Similarly, Supplementary Figure 3 illustrates the effect of the number of contacts. Figure 5 illustrates how the distance from the catalytic effect predictable contributions to the fitness, and solubility. Interestingly the two outcomes show an opposite dependence on the distance from the catalytic site, as was seen using a classification model [20].

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

Validation of Correctness- an important consequence of explaining the effects is that by validating that the factors that one believes are indeed the most relevant ones in calculations, a sense of correctness of individual predictions may be developed.

Protein Engineering - The analyses such as understanding the tradeoffs between solubility and fitness [19] have been strongly motivated from the perspective of designing better proteins. The same is true when a direct correlation between measured fitness and solubility cannot be inferred (Supplementary Figure 4), where the individual components are more predictable and hence reliable.

Developing intuitions – The developments in the fields of deep mutational scan have a great value to add to the intuitions that are pedagogically taught, such as larger conservation implies greater impact. The patterns of dependence of the mutational effects on the individual parameters allows one to go beyond predictions to learning and developing of rules.

Understanding protein function is not easy. At times a single mutation can lead to deleterious effects, and yet evolutionarily one sees homologous proteins with as little as 50% sequence identity performing similar functions. One has to resort to advanced AI methods to predict the effects of one or more mutations, to note how

mutations affect, or compensate for each other, or to reduce the experimentation required. Introducing explainability into the analyses can potentially help in an improved learning about how mutations affect the proteins.

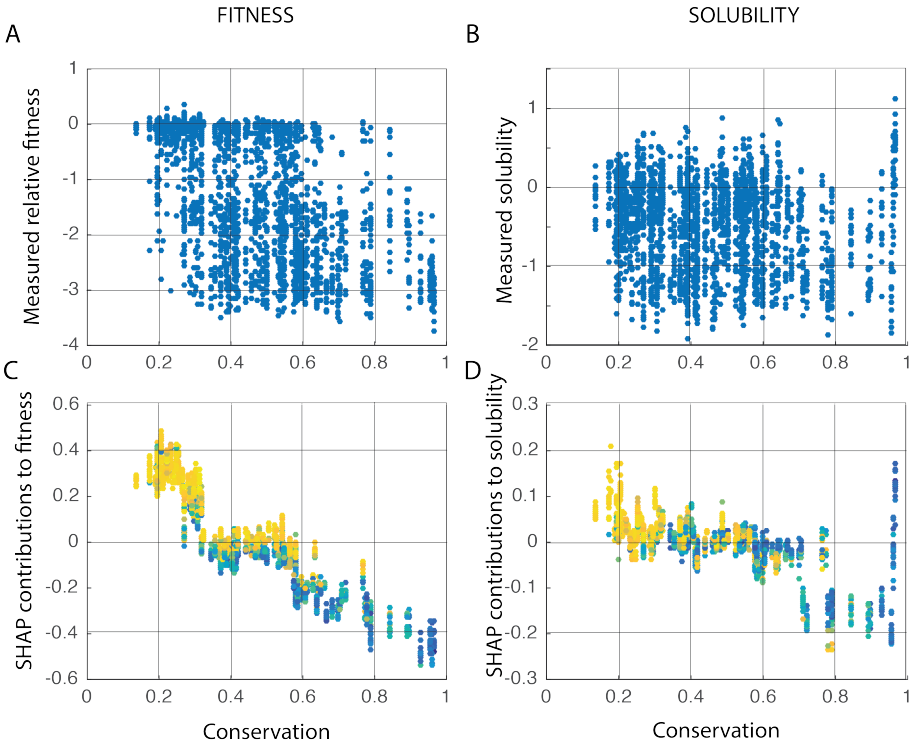


Figure 4. Extracting relations. The scatter plots of **A.** fitness and **B.** solubility relative to conservation do not show a clear pattern of what one can expect from substituting an amino acid with high conservation. On the contrary, the SHAP contributions to **C.** fitness and **D.** solubility show a very clear pattern of reducing SHAP values with increasing conservation which suggests that the fitness and solubility decrease with the substitution of a conserved amino acid. The colorbar is the same as in Figure 3, and represents the observed fitness changes.

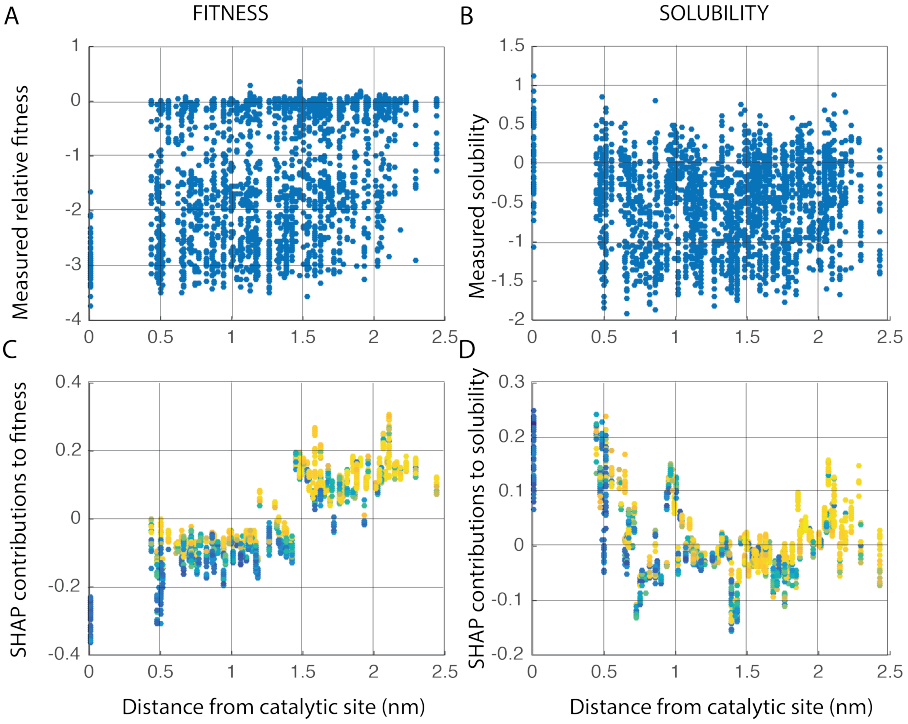


Figure 5. Extracting relations. The scatter plots of **A.** fitness and **B.** solubility relative to conservation do not show a clear pattern of what one can expect from substituting an amino acid which is further away from the catalytic

site. On the contrary, the SHAP contributions to **C.** fitness and **D.** solubility show a very clear pattern. Interestingly the effect of the distance from the catalytic is opposite on fitness and solubility. This quantitative trend is in line with the observations from the classification model [Klesmith et al. 2016]. The colorbar is the same as in Figure 3, and represents the observed fitness changes.

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

References.

1. Crow JF (2000) The origins, patterns and implications of human spontaneous mutation, Nat. Rev. Genetics, 1: 40–47. doi:10.1038/35049558

2. Stewart PS, Costerton JW (2001) Antibiotic restistance of bacteria in biofilms, Lancet, 358, 135-138.

3. Fersht AR (1987) Dissection of the structure and activity of the tyrosyl-tRNA synthetase by site-directed mutagenesis, Biochemistry, 26: 8031-8037.

4. Cunningham BC, Wells JA (1989), High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis, Science, 244: 1081-1085.

5. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, Fields S (2010) High-resolution mapping of protein sequence-function relationships. Nat Methods 7: 741–746.

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

7. Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, Fowler DM, Parvin JD, Shendure J, Fields S (2015) Massively parallel functional analysis of BRCA1 RING domain variants. Genetics 200: 413–422.

8. Weile, J. et al. (2017) A framework for exhaustively mapping functional missense variants. Mol. Syst. Biol. 13: 957.

9. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, Marks DS (2017) Mutation effects predicted from sequence co-variation, *Nat. Biotechnology,* 35, 128-135.

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

11. Hecht M, Bromberg Y, Rost B (2015) Better prediction of functional effects for sequence variants, BMC Genomics. 2015; 16(Suppl 8): S1.

12. Ribeiro MT, Singh S, Guestrin, C (2016) Why should I trust you?: Explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, 1135–1144.

13. Datta A, Sen S, Zick, Y (2016) Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems, Security and Privacy (SP), 2016 IEEE Symposium on. IEEE.

14. Štrumbelj E, Kononenko I (2013) Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems, 41(3), 647–665. doi:10.1007/s10115-013-0679-x

15. Breiman L (2001) Statistical Modeling: The Two Cultures, Statistical Science, 16, 199-215.

16. Lundberg SM, Lee, S-I (2017) A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems.

17. Lundberg SM, Erion GG, Lee S-I (2018) Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

18. Shapley LS (1953) A value for n-person games, vol II of Contributions to the theory of games. Princeton University Press, Princeton.

19. Stiffler MA, Hekstra DR, Ranganathan R (2015) Evolvability as a function of purifying selection in TEM-1 β-Lactamase, Cell 160: 882–892.

20. Klesmith JR, Bacik J-P, Wrenbeck EE, Michalczyk R, Whitehead TA (2017) Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning, Proc. Natl. Acad. Sci. USA, 114, 2265-2270.

21. Minasov G, Wang X, Shoichet BK (2002) An ultrahigh resolution structure of TEM-1 beta-lactamase suggests a role for Glu166 as the general base in acylation, J. Am.Chem.Soc. 124: 5333-5340.

22. Sruthi CK, Prakash MK (2017) Deep2Full: Predictive model for complementing phenotypic outcomes in a deep mutational scan using protein sequence and structure information, biorxiv, https://doi.org/10.1101/217158

23. Chennubhotla C, Bahar I (2007) "Signal propagation in proteins and relation to equilibrium fluctuations," PLoS Comp. Bio., 3, 1716–1726.

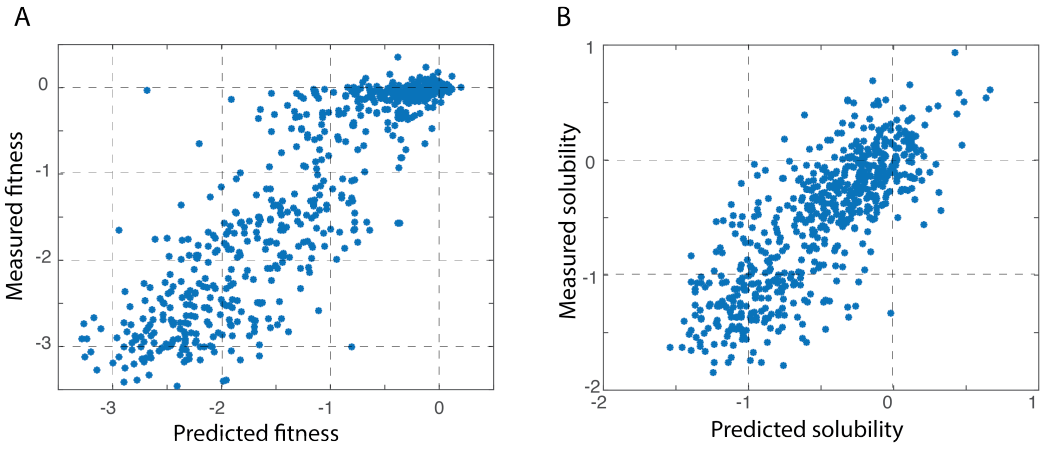
24. Henikoff S, Henikoff J (1992) Amino-acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. USA, 89, pp. 10915–10919.

25. Sruthi CK, Prakash MK (2018) Amino acid impact factor, PLoS ONE 13(6): e0198645. https://doi.org/10.1371/journal.pone.0198645.

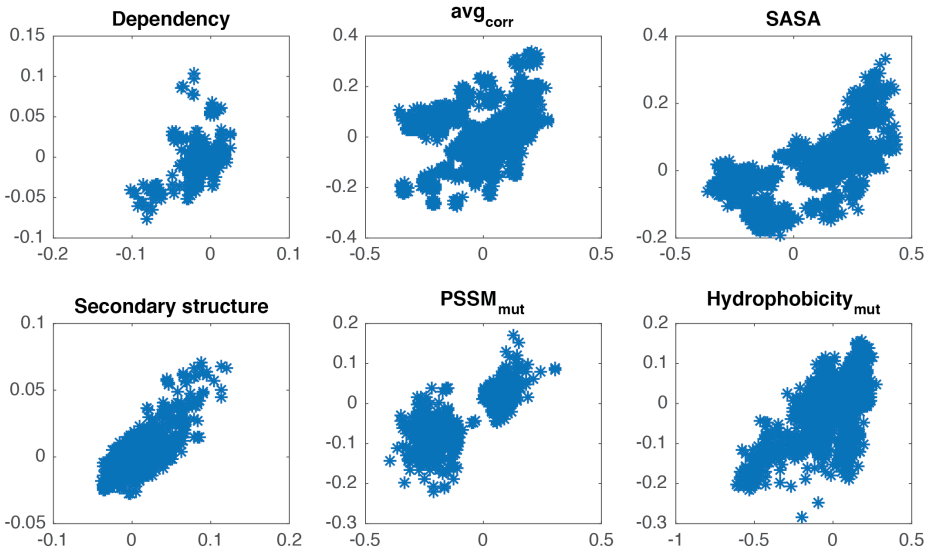
Supplementary Information

Interpreting each individual mutational effect prediction

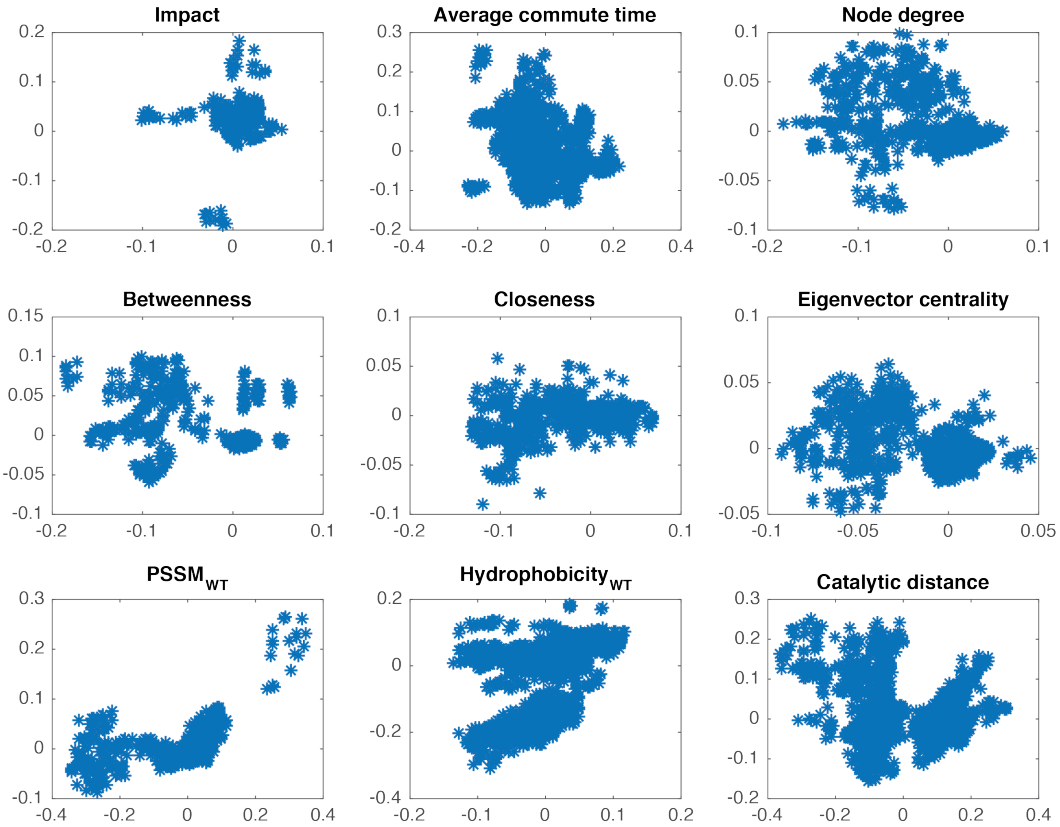
C. K. Sruthi and Meher K. Prakash*



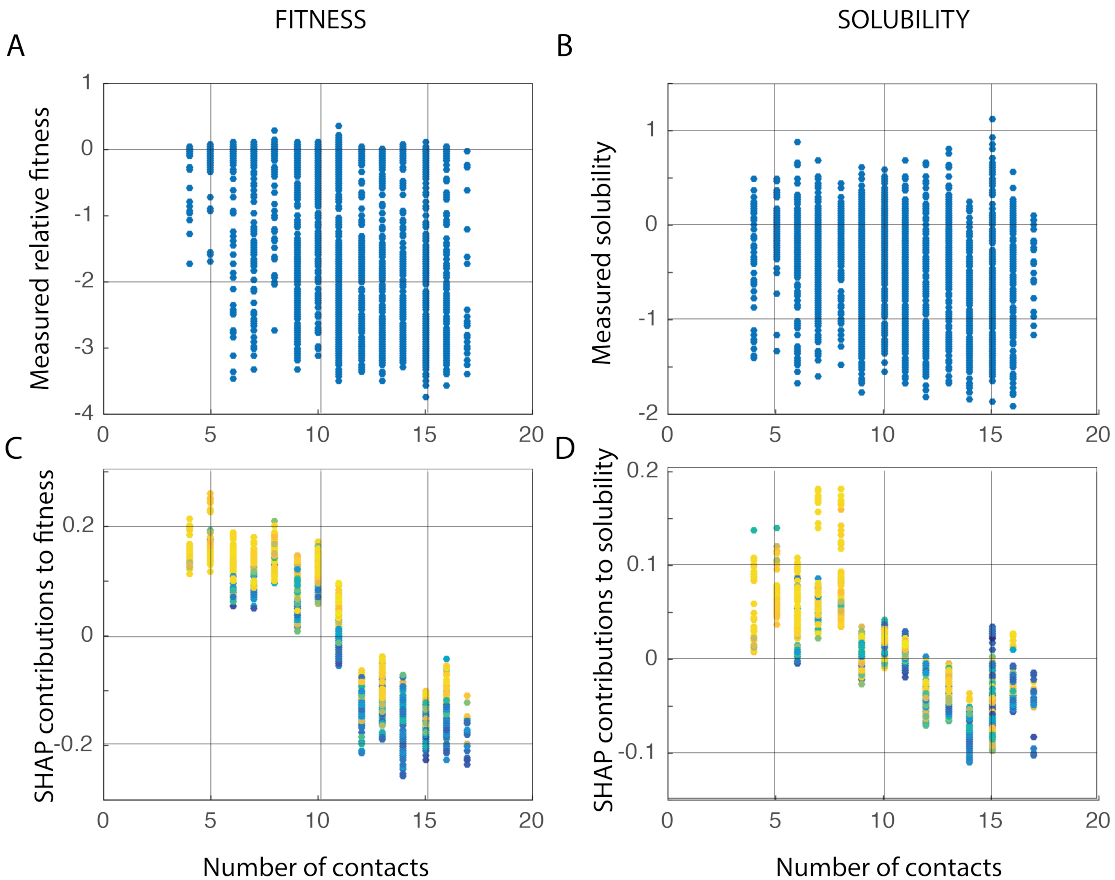
Supplementary Figure 1. Quality of predictions. 75% of the mutational data was used for training the models. The quality of the model was judged by comparing the predictions to the observations. The scatter plots show that both **A.** fitness and **B.** solubility can be predicted with a good reliability.



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

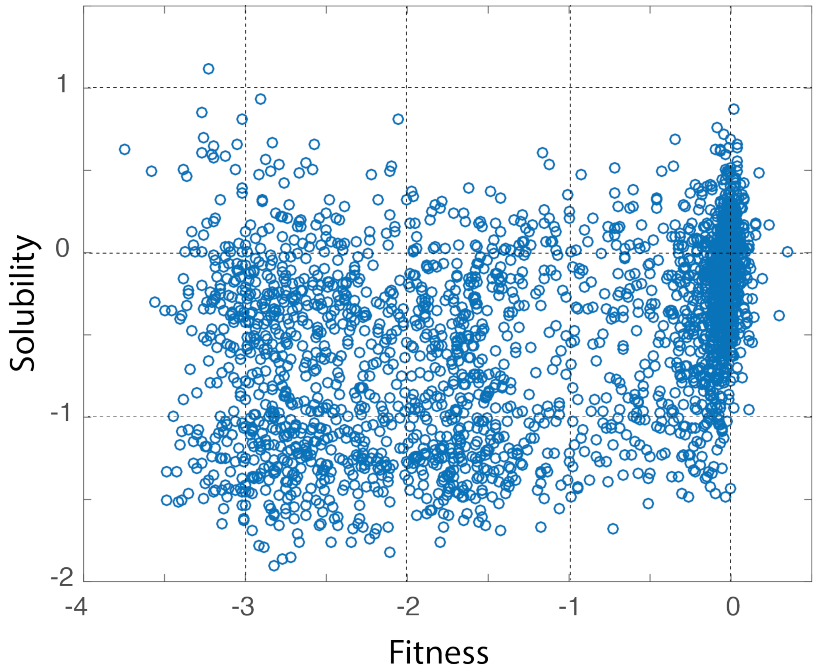


Supplementary Figure 2B. Correlation of contributions to solubility and fitness. A continuation Supplementary Figure 2A, showing the contributions from some more parameters to fitness and solubility are poorly correlated.



Supplementary Figure 3. Extracting relations. The scatter plots of **A.** fitness and **B.** solubility relative to the number of contacts the wild type amino acid has in its structure do not show a clear pattern. On the contrary, the

SHAP contributions to **C.** fitness and **D.** solubility show a very clear pattern of reducing SHAP values with increasing contacts, which suggests that the fitness and solubility decrease with the substitution of a tightly packed amino acid. The colorbar is the same as in **Figure 3**, and represents the observed fitness changes.



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