Amino acid exchangeability and surface accessibility underpin the effects of single substitutions

Berk A. Alpay[1*], Piyush Nanda[2,3], Emma Nagy[3], Michael M. Desai[4,5]

Deep mutational scans have measured the effects of many mutations on many different proteins. Here we use a collection of such scans to perform a statistical meta-analysis of the effects of single amino acid substitutions. Specifically, we model the relative deleteriousness of each substitution in each deep mutational scan with respect to the identities of the wildtype and mutant residues, and the wildtype residue's surface accessibility. This model explains much of the variance in mutational effects and quantifies physicochemical trends underlying them, including by yielding an empirical amino acid exchangeability matrix.

It is of fundamental interest in medicine [1], molecular evolution [2], and protein engineering [3] to understand the effects of mutations on proteins. However, proteins are diverse [4] and they and their properties are biophysically complex [5]. Accordingly, experimental techniques, most prominently deep mutational scanning [6], have been developed to measure the effects of mutations at high throughput on proteins of interest. Different protein properties can be assayed in this way, common effects measured being those on expression [7], binding affinity to a ligand [8], and catalytic activity [7, 9].

Machine learning models have been developed to predict the deleteriousness of mutations by comparing the mutated sequence to natural sequences that have been recorded in large repositories [16]. In rendering their predictions, these models have traditionally required input of sequences homologous to the protein of interest [12, 17, 18]. Protein language models [19, 20], recent alternatives, instead learn from a large and general corpus of protein sequences before tailoring their predictions to the given protein. Predictions of machine learning models are typically compared against deep mutational scans for validation. Databases such as MaveDB [21] and ProteinGym [15] are useful to this end because they compile deep mutational scans in a standard format. ProteinGym is designed for benchmarking models and thus also provides the predictions of several dozen machine learning models alongside experimentally measured effects.

Even since before deep mutational scanning and despite the biophysical idiosyncrasies of protein function, trends in mutational effects have been noted. Work with individual proteins revealed structural considerations in substitution tolerance [22]. It is well-recognized for example that polar substitutions to buried residues tend to be disruptive [23, 24], more so than substitutions to solvent-accessible residues [25, 26]. Similarly, substitutions to loops and turns are typically better tolerated than those to alpha helices and beta sheets [26], with mutations to proline posing a particular danger to these structured regions [27]. Indeed, methods of predicting mutational effects with respect to human disease have often incorporated physicochemical features [28–31], typically as part of a collection of information including sequence alignments.

Deep mutational scanning data now enables us to understand these physicochemical trends more precisely and with greater statistical power, an opportunity others have taken to deliver insightful analyses [32, 33]. With this data Gray et al. for example argued histidine and asparagine substitutions are the ones most demonstrative of the effects of other substi-

* berk_alpay@g.harvard.edu
[1]Department of Systems Biology, Harvard Medical School
[2]Biological and Biomedical Sciences Program, Harvard University
[3]Department of Molecular and Cellular Biology, Harvard University
[4]Department of Organismic and Evolutionary Biology, Harvard University
[5]Department of Physics, Harvard University

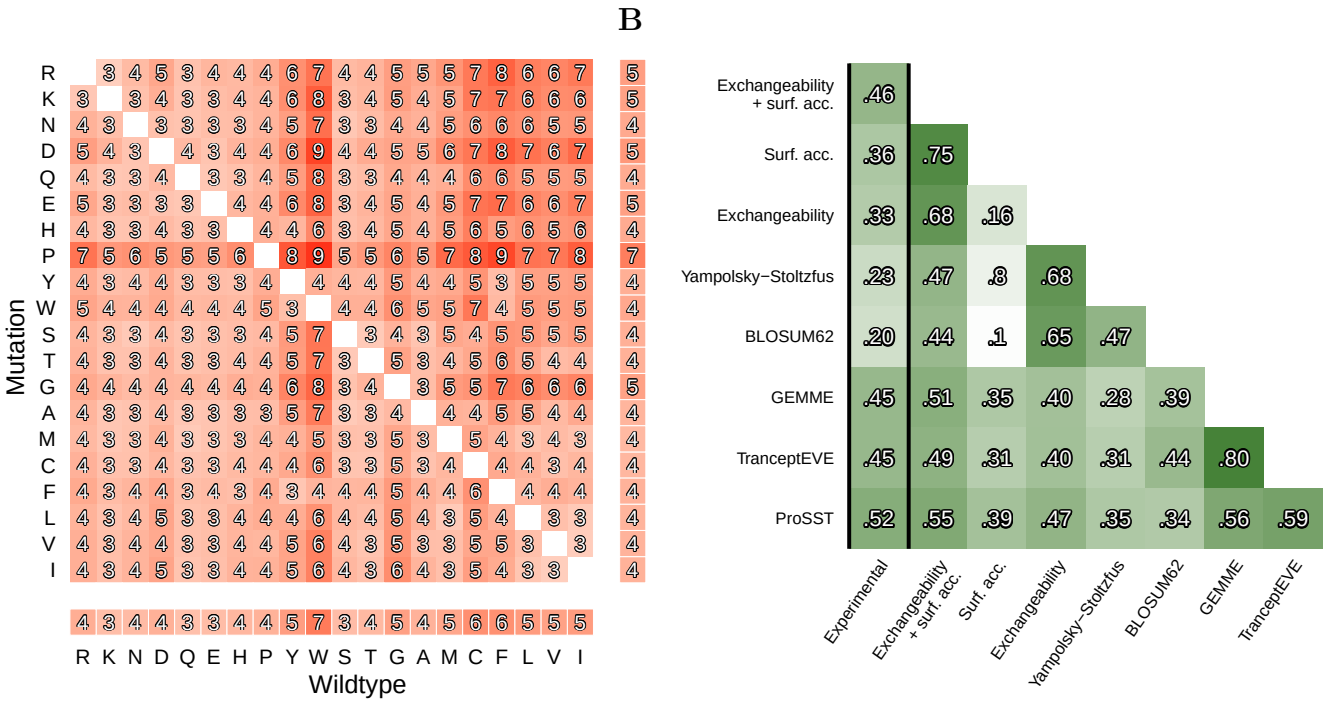


Figure 1: (A) Amino acid exchangeabilities as inferred from a model of ranks of mutational effects within deep mutational scans. Scores in the margins represent averages for the wildtype or mutant residues. Higher scores indicate greater deleteriousness, with scores rescaled $\times 10$ and rounded to the nearest integer for compactness. **(B)** The average per-scan Spearman rank correlation of experimental single-substitution effects, model predictions, and measures of amino acid exchangeability. Predictions of our exchangeability and combined exchangeability and surface accessibility models are from ten-fold cross-validation among proteins. Included are the Yampolsky-Stoltzfus exchangeability matrix [10] and BLOSUM62 [11]. The sequence-based machine learning models are GEMME [12], an alignment-based model, TranceptEVE [13], an alignment-based and protein language model, and ProSST [14], a structure and protein language model [15].

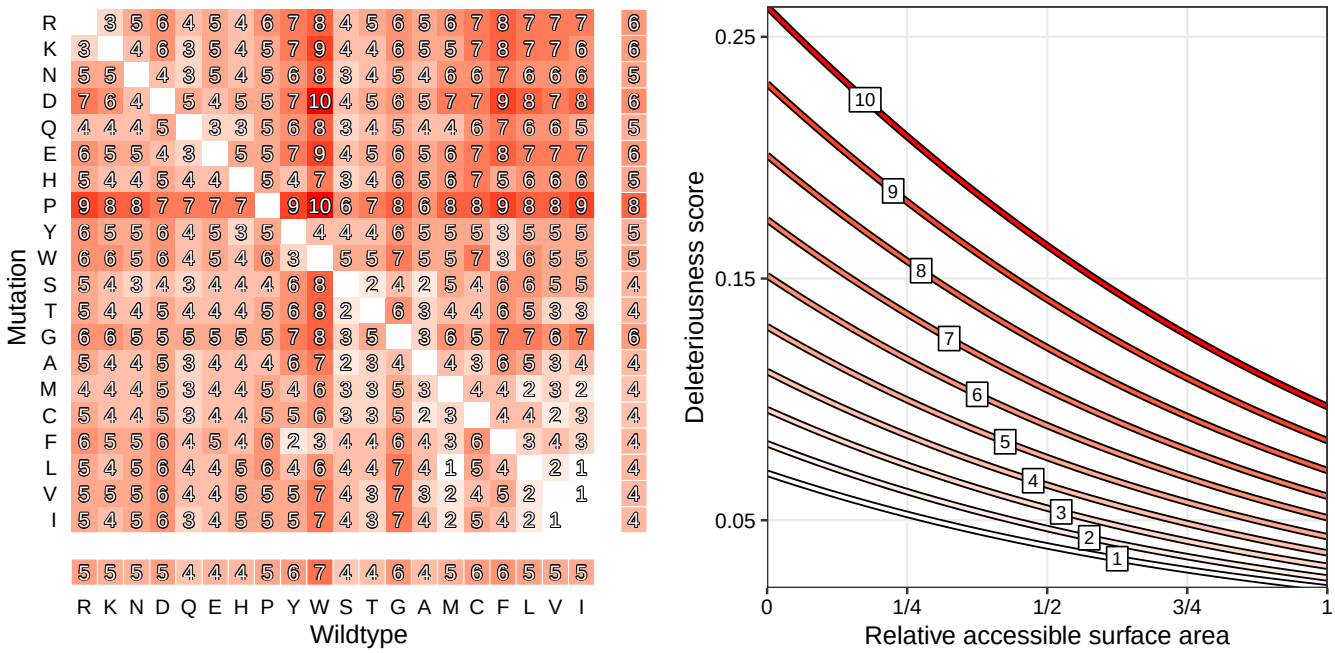
tutions [34] as opposed to alanine scanning [35], and Dunham and Beltrao analyzed dozens of deep mutational scans, categorizing protein sites by patterns of mutational effects and interpreting these categories with respect to local structural context [36]. Mutational effects have been modeled with empirical amino acid exchangeability matrices [37, 38], and with the surface accessibility of the substituted residue [39, 40]. Schulze and Lindorff-Larsen combined these features to infer a separate exchangeability matrix for buried and exposed residues [41] in analyzing effects of mutations on the protein's cellular abundance.

In this study, we measure aggregate physicochemical commonalities among mutational effects and quantify the degree to which they explain the large and varied collection of deep mutational scanning data now at our disposal. We use a statistical rank model to analyze these various deep mutational scans in concert, first with respect only to aggregate amino acid exchangeability, and then with respect also to surface accessibility. We analyze these physicochemical inferences and show that this model, on par with sequence-based machine learning models, explains much of the variance in the experimentally measured effects of single substitutions, and does so in terms of only an amino acid exchangeability matrix and a parameter for the effect of surface accessibility.

Results

We use the deep mutational scans that have been compiled in ProteinGym, which is composed of a variety of more than 200 assays (Figure S1) and nearly 700,000 single amino acid substitutions, which are well-sampled at the level of sites; for most sites in the data, the effects of substituting the wildtype with each of the 19 other proteinogenic amino acids have

A



B

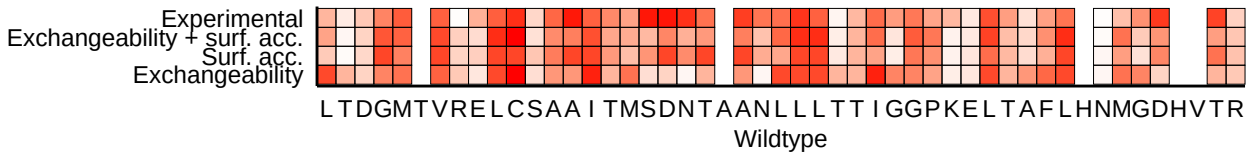


Figure 2: (A) Deleteriousness scores with respect to the amino acids being exchanged and the wildtype residue's surface accessibility. For illustration, amino acid scores are binned from one through ten, and the deleteriousness curve with respect to surface accessibility is shown using the central score of each bin. **(B)** A demonstration of predicting the ranked deleteriousness of histidine substitutions at residues 111 to 159 of beta-lactamase based on amino acid exchangeability, surface accessibility, and both, compared to the measurements of a deep mutational scan by Firnberg et al. [42]. Some loci lack experimental measurements. The Spearman correlation of the full model's predictions with the experiment is 0.68 in this region.

been measured (Figure S2A). Although effect scores are ordered by deleteriousness within scans, due to differences in methodology they can follow widely varying scales between scans [36, 43]. Previous methods have accounted for this problem by restricting scans to those that use a specific experimental technique [41] or applying a transformation to scores toward a standard scale [36]. Deep mutational scanning scores are however consistently ordered by deleteriousness, a property which has been exploited by methods that normalize the ranks of scores within scans [37]. To better account for potential differences in the kinds of sites represented between scans, and therefore inconsistencies even in such normalized scores, we explicitly model the ranks of the effects by deleteriousness within each scan, conditioning on features of the site.

Our models are effectively Plackett-Luce models [44, 45], in that each substitution has a corresponding deleteriousness score which we infer, and in that the probability of an ordering of substitutions (by decreasing experimentally measured deleteriousness) with corresponding scores $\theta_1, \theta_2, \ldots, \theta_n$ is $\prod_{i=1}^{n}(\theta_i / \sum_{j=i}^{n} \theta_j)$. The latent deleteriousness scores are on a consistent scale: from zero to one, the latter representing maximal deleteriousness. (Although beneficial effects can be detected by deep mutational scans, we assume the best effect that can be expected is total neutrality.) In contrast to the canonical Plackett-Luce model, our

models infer the θ by regression, analyzing the rank of each substitution in each deep mutational scan with respect to the identities of the amino acids being exchanged, and in our full model, the wildtype residue's surface accessibility.

We first applied a model that determines the score of a particular substitution purely with respect to the identities of the amino acids being exchanged. Let w_i represent the identity of the wildtype amino acid of the ith most deleterious substitution, and let m_i represent that of the mutant. Then in this model the deleteriousness score of the substitution is β_{w_i,m_i}, so that $\theta_i = \beta_{w_i,m_i}$. We set a prior that the set of $20 \times 19 = 380$ deleteriousness scores, for every possible exchange of amino acids, are independent and identically distributed as standard normal random variables before each being standard-logistically transformed to fall between zero and one.

Under this model, we inferred from the collected deep mutational scans the aggregate deleteriousness of substituting each amino acid for each other (Figure 1A). These scores are learned directly from substitutions in proteins that are shaped by evolution and are in at least some sense functional. For example, there are several-fold more leucine-wildtype sites in the data than cysteine-wildtype sites (Figure S2B). The inferences implicitly reflect other physicochemical trends than just the properties of the amino acids being exchanged. The relative frequencies of amino acids vary across surface accessibility (Figure S3) and secondary structure. For example, based on its tendency to disrupt secondary structure, one should expect that the typical wildtype proline is not central in alpha helices nor beta sheets, and based on their hydrophobicities, that the surface accessibility of the typical wildtype tryptophan is less than that of the typical wildtype lysine. The exchangeabilities we infer implicitly condition on these biases.

Altogether, these amino acid exchangeabilities explain a considerable proportion of the variance in the experimentally ranked effects (Figure 1B) and reflect known physicochemical relationships. The three least deleterious scores are assigned to the substitution of arginine for lysine followed by isoleucine for valine and valine for isoleucine. The five wildtype amino acids most prone to deleterious substitution are tryptophan followed by cysteine, phenylalanine, leucine, and isoleucine, which, excepting cysteine, are relatively large, while moderately sized amino acids are more tolerant (Figure S4). Cysteine's obstinacy as a wildtype may be due to its unique ability to form disulfide bonds and its potential to facilitate catalytic interactions.

There are prominent asymmetries in exchangeability, in that a substitution can be more deleterious in one direction than the reverse (Figure S5) [37]. The mutant amino acid that is on average the most deleterious is proline, but proline-wildtype sites are not particularly intolerant. This contrast reflects proline's tendency to disrupt secondary structure, while proline-wildtype sites have already in a sense accepted this biophysical limitation, just as a mutant cysteine residue would not be expected to form a disulfide bridge. The primary asymmetry is that substituting uncharged amino acids with charged ones is generally particularly deleterious, while substituting charged amino acids with uncharged ones is less so, possibly reflecting that uncharged amino acids, which are often buried in the hydrophobic core, are particularly intolerant to hydrophilic substitutes.

ProteinGym categorizes scans into broad types of protein properties that they assess: catalytic or biochemical activity, binding affinity to a target, expression, organismal growth rate, and stability. We inferred exchangeabilities from training on each of these categories separately and observed that they are fairly robust to the types of assays used to train them, with some deviations (Figure S6). Notably, replacing hydrophilic amino acids with hydrophobic ones tends to be more deleterious than usual among binding assays.

We next applied a separate model that includes surface accessibility in addition to amino acid exchangeability. In this model, the deleteriousness score of each substitution remains between zero and one, but it is computed as a logistic regression on the identities of the wildtype and mutant amino acids as well as the wildtype surface accessibility. The empirical trends of deleteriousness with respect to surface accessibility (Figure S3) indicate that this relationship is fairly consistent across substitutions, so we model it with a single parameter β_{SA}. Specifically, we set each score $\theta_i = \mathrm{logistic}(\beta_{w_i,m_i} + \beta_{\mathrm{SA}} s_i)$, where s_i is the AlphaFold2-predicted wildtype surface accessibility at the site of the ith most deleterious substitution. We set an independent standard normal prior for each β_{w_i,m_i}, and for β_{SA}. The inferred model is shown in Figure 2A. The inferences align fairly well with the empirical trends, and the exchangeability ranks, inde-

pendent of surface accessibility, are very similar (with Spearman correlation 0.91) to those inferred by the purely amino acid-based model. Figure 2B illustrates an example of the model's predictions.

Surface accessibility alone explains the measured ranks better than exchangeability alone, and including both considerably increases the variance explained to a degree on par with sequence-based machine learning models (Figure 1B). (We also computed the average per-scan Spearman correlation of MIF [46], an inverse folding model, with deep mutation scans as 0.41, which is comparable to the others.) There is generally some but not overwhelming agreement about predictions among our models and the sequence-based models; the Spearman correlation between the predictions of our exchangeability matrix and those of TranceptEVE, for example, is 0.40. The Spearman correlation of our exchangeability and surface accessibility model with experimental measurements varies fairly widely across deep mutational scans but is more consistent among larger scans, for which the Spearman correlation ranges between approximately 0.3 and 0.6. The model is fairly similar in its average per-scan Spearman correlation across types of scans: activity (0.43), binding (0.39), expression (0.44), organismal fitness (0.39), and stability (0.57) (Figure S7).

Discussion

By building statistical models of deep mutational scans, we collectively analyzed the relative deleteriousness of single amino acid substitutions aggregated across different proteins and protein properties with respect to amino acid exchangeability and surface accessibility. We found that these features together, on par with the predictions of even sequence-based machine learning models that involve many more parameters, tend to explain much of the variance in deleteriousness ranks within deep mutational scans.

Models that combine site-wise evolutionary conservation with amino acid exchangeability [37], and site-wise conservation with solvent accessibility [40] are also relatively simple and explain a good deal of mutational effects across a variety of types of assays. Unlike these models, ours do not condition on homologous sequence data and may overlook, for example, surface-accessible functional sites. Instead, our model essentially provides a prior over mutational effects given aggregate patterns, one we find is strong and which we anticipate will be tangibly useful toward various ends, perhaps as a prior for parameters of machine learning models especially when homologous sequence data is limited, as a baseline [38] for models that issue predictions more strongly tailored to the protein of interest, and for design, such as when a variety of still-functional variants are desired as starting points in engineering a protein. Empirical departures from the prior could also be used as evidence of functionally significant residues [41, 47].

We compared our models' scores to the predictions of sequence-based models to understand how well our models explain the relative deleteriousness of single amino acid substitutions. These results should not be taken to mean that our aggregate physicochemical models are overall superior to sequence-based models: sequence-based models can perform tasks, including structure prediction [48] and sequence generation [20, 49], of which our models are incapable; they are unsupervised and trained on orthogonal data while ours were trained directly on deep mutational scans; and our models analyze effects of only single substitutions and therefore cannot capture epistasis, which has been observed to pervade genotype-phenotype relationships in proteins [50, 51] and which sequence models are often designed to capture [12, 18].

An immediate product of our statistical analysis was an empirical amino acid exchangeability matrix which quantifies the relative deleteriousness of each of the 380 possible amino acid substitutions. The matrix is asymmetric, in that the deleteriousness score depends not only on which amino acids are being swapped, but on which is the wildtype and which the mutant. It therefore implicitly takes into account typical physicochemical features and roles of amino acids. This matrix contrasts with substitution matrices based on evolutionary conservation, e.g. BLOSUM [11] and PAM [52], and physicochemical distance matrices [53–55], which are symmetric and do not consider mutational effects directly; substitution matrices implicitly reflect factors like codon bias and the structure of the genetic code [10] and are not intended for mutation effect prediction [17] while physicochemical distance matrices represent differences between amino acids but not directly with respect to their roles in proteins.

Yampolsky and Stoltzfus published an empirical exchangeability matrix in 2005 using data from systematic exchange studies [10], precursors to deep mutational scans. We can now draw from several times more proteins and substitutions [37]. Our matrix

has a 0.64 Spearman correlation with the Yampolsky-Stoltzfus matrix and yields somewhat similar effect predictions as both this matrix and BLOSUM62 while ultimately explaining more variance in the empirical ranks (Figure 1B). We recover similar patterns in exchangeability as recent studies: our matrix has a Spearman correlation of 0.94 with one by Munro and Singh [37] and one by Høie et al. [38].

Our physicochemical models explain a substantial proportion of the variance in the relative deleteriousness of mutations without accounting for the protein properties being assayed, and the inferences from training on assay types separately are similar to those learned from the whole data. Indeed, we find that not only do surface accessibility and measures of amino acid exchangeability explain much of the variance in effects on expression [41] and stability [56], they also do well in explaining effects on other aspects of protein function. And although almost all (64 of 66) of the stability scans are relatively small and come from a single study [57] that introduced a technique for scanning domains at scale (the next-most prolific study is associated with only four scans), they are the best-explained by exchangeability and surface accessibility. These results are consistent with there being intrinsic properties of proteins, including stability [3, 32] but also perhaps collateral cellular effects such as those on gene expression and pleiotropic interactions [33, 58], that underlie assay measurements, and therefore that substitutions might, in impacting these properties, have comparable effects on higher-order properties such as binding affinity and catalytic activity.

We have shown that aggregate amino acid exchangeability and a surface accessibility parameter explain much of the effects of single substitutions across various proteins and types of deep mutational scans. Other physicochemical features could be incorporated, although there are natural limits on how much variance in deep mutational scans can be explained given sequencing [59] and experimental noise, which can lead to divergent effect scores even between experimental replicates [60]. Idiosyncratic interactions between sites [51], the intricacies of protein function, and multi-site substitutions [61] add to the complexity of intuiting the effects of mutations. However, physicochemical trends such as those we have analyzed support that strong prior understanding of effects at short mutational distance can be formed, and may aid in the challenge to understand the grammar of proteins.

Methods

We extracted surface accessibility data by using the software DSSP [62, 63] on the AlphaFold2-predicted [64] structures of each protein as provided by ProteinGym [15] version 1.2. We also used the metadata, model predictions, and experimental measurements for 217 deep mutational scans from this version of ProteinGym. We performed maximum a posteriori estimation of model parameters using Stan [65]. Amino acid volumes are the consensus values calculated by Perkins [66], and amino acids are ordered in figures according to Kyte and Doolittle's hydropathy index [67]. Code can be accessed at https://github.com/berkalpay/metamutation.

Acknowledgements

Thank you to Pascal Notin for useful discussions about ProteinGym, and to Wendy Valencia-Montoya for her feedback on the text. B.A.A. was supported by an NSF Graduate Research Fellowship under Grant No. DGE-2140743.

References

1. Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R. & Alexov, E. Molecular mechanisms of disease-causing missense mutations. *Journal of Molecular Biology* 425, 3919–3936 (2013).
2. Maynard Smith, J. Natural selection and the concept of a protein space. *Nature* 225, 563–564 (1970).
3. Bloom, J. D. & Arnold, F. H. In the light of directed evolution: pathways of adaptive protein evolution. *Proceedings of the National Academy of Sciences* 106, 9995–10000 (2009).
4. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* 28, 235–242 (2000).
5. Whisstock, J. C. & Lesk, A. M. Prediction of protein function from protein sequence and structure. *Quarterly Reviews of Biophysics* 36, 307–340 (2003).
6. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nature Methods* 11, 801–807 (2014).
7. Vanella, R. *et al.* Understanding activity-stability tradeoffs in biocatalysts by enzyme proximity sequencing. *Nature Communications* 15, 1807 (2024).

8. Starr, T. N. *et al.* Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* 182, 1295–1310 (2020).
9. Jacquier, H. *et al.* Capturing the mutational landscape of the beta-lactamase TEM-1. *Proceedings of the National Academy of Sciences* 110, 13067–13072 (2013).
10. Yampolsky, L. Y. & Stoltzfus, A. The exchangeability of amino acids in proteins. *Genetics* 170, 1459–1472 (2005).
11. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* 89, 10915–10919 (1992).
12. Laine, E., Karami, Y. & Carbone, A. GEMME: a simple and fast global epistatic model predicting mutational effects. *Molecular Biology and Evolution* 36, 2604–2619 (2019).
13. Notin, P. *et al.* TranceptEVE: Combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. *bioRxiv,* 2022–12 (2022).
14. Li, M. *et al.* ProSST: Protein language modeling with quantized structure and disentangled attention. *Advances in Neural Information Processing Systems* 37, 35700–35726 (2025).
15. Notin, P. *et al.* ProteinGym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems* 36 (2024).
16. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* 45, D158–D169 (2017).
17. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Research* 11, 863–874 (2001).
18. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nature Biotechnology* 35, 128–135 (2017).
19. Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems* 34, 29287–29303 (2021).
20. Su, J. *et al.* SaProt: Protein language modeling with structure-aware vocabulary. *bioRxiv,* 2023–10 (2023).
21. Esposito, D. *et al.* MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biology* 20, 1–11 (2019).
22. Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science* 247, 1306–1310 (1990).
23. Perutz, M., Kendrew, J. & Watson, H. Structure and function of haemoglobin: II. Some relations between polypeptide chain configuration and amino acid sequence. *Journal of Molecular Biology* 13, 669–678 (1965).
24. Reidhaar-Olson, J. F. & Sauer, R. T. Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. *Science* 241, 53–57 (1988).
25. Schwehm, J. M., Kristyanne, E. S., Biggers, C. C. & Stites, W. E. Stability effects of increasing the hydrophobicity of solvent-exposed side chains in staphylococcal nuclease. *Biochemistry* 37, 6939–6948 (1998).
26. Guo, H. H., Choe, J. & Loeb, L. A. Protein tolerance to random amino acid change. *Proceedings of the National Academy of Sciences* 101, 9205–9210 (2004).
27. Chou, P. Y. & Fasman, G. D. Prediction of protein conformation. *Biochemistry* 13, 222–245 (1974).
28. Chasman, D. & Adams, R. M. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *Journal of Molecular Biology* 307, 683–706 (2001).
29. Sunyaev, S. *et al.* Prediction of deleterious human alleles. *Human Molecular Genetics* 10, 591–597 (2001).
30. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* 7, 248–249 (2010).
31. Ponzoni, L., Peñaherrera, D. A., Oltvai, Z. N. & Bahar, I. Rhapsody: predicting the pathogenicity of human missense variants. *Bioinformatics* 36, 3084–3092 (2020).
32. Stein, A., Fowler, D. M., Hartmann-Petersen, R. & Lindorff-Larsen, K. Biophysical and mechanistic models for disease-causing protein variants. *Trends in Biochemical Sciences* 44, 575–588 (2019).
33. Cagiada, M. *et al.* Understanding the origins of loss of protein function by analyzing the effects of thousands of variants on activity and abundance.

Molecular Biology and Evolution 38, 3235–3246 (2021).
34. Gray, V. E., Hause, R. J. & Fowler, D. M. Analysis of large-scale mutagenesis data to assess the impact of single amino acid substitutions. *Genetics* 207, 53–61 (2017).
35. Cunningham, B. C. & Wells, J. A. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science* 244, 1081–1085 (1989).
36. Dunham, A. S. & Beltrao, P. Exploring amino acid functions in a deep mutational landscape. *Molecular Systems Biology* 17, e10305 (2021).
37. Munro, D. & Singh, M. DeMaSk: a deep mutational scanning substitution matrix and its use for variant impact prediction. *Bioinformatics* 36, 5322–5329 (2020).
38. Høie, M. H., Cagiada, M., Frederiksen, A. H. B., Stein, A. & Lindorff-Larsen, K. Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell Reports* 38, 110207 (2022).
39. Sruthi, C. K., Balaram, H. & Prakash, M. K. Toward developing intuitive rules for protein variant effect prediction using deep mutational scanning data. *ACS Omega* 5, 29667–29677 (2020).
40. Tsishyn, M., Hermans, P., Pucci, F. & Rooman, M. Residue conservation and solvent accessibility are (almost) all you need for predicting mutational effects in proteins. *Bioinformatics,* btaf322 (2025).
41. Schulze, T. K. & Lindorff-Larsen, K. Effects of residue substitutions on the cellular abundance of proteins. *eLife* 13, RP103721 (2024).
42. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A comprehensive, high-resolution map of a gene's fitness landscape. *Molecular Biology and Evolution* 31, 1581–1592 (2014).
43. Schulze, T. K., Blaabjerg, L. M., Cagiada, M. & Lindorff-Larsen, K. Supervised learning of protein variant effects across large-scale mutagenesis datasets. *bioRxiv,* 2025–04 (2025).
44. Luce, R. D. *Individual Choice Behavior* (Wiley New York, 1959).
45. Plackett, R. L. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics* 24, 193–202 (1975).
46. Yang, K. K., Zanichelli, N. & Yeh, H. Masked inverse folding with sequence transfer for protein representation learning. *Protein Engineering, Design and Selection* 36, gzad015 (2023).
47. Rao, J. *et al.* Rosace-AA: Enhancing Interpretation of Deep Mutational Scanning Data with Amino Acid Substitution and Position-Specific Insights. *bioRxiv,* 2025–01 (2025).
48. Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PloS One* 6, e28766 (2011).
49. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods* 15, 816–822 (2018).
50. Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. *Protein Science* 25, 1204–1218 (2016).
51. Buda, K., Miton, C. M. & Tokuriki, N. Pervasive epistasis exposes intramolecular networks in adaptive enzyme evolution. *Nature Communications* 14, 8508 (2023).
52. Dayhoff, M., Schwartz, R. & Orcutt, B. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure* 5, 345–352 (1978).
53. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* 185, 862–864 (1974).
54. Miyata, T., Miyazawa, S. & Yasunaga, T. Two types of amino acid substitutions in protein evolution. *Journal of Molecular Evolution* 12, 219–236 (1979).
55. Kawashima, S. *et al.* AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research* 36, D202–D205 (2007).
56. Caldararu, O., Blundell, T. L. & Kepp, K. P. Three simple properties explain protein stability change upon mutation. *Journal of Chemical Information and Modeling* 61, 1981–1988 (2021).
57. Tsuboyama, K. *et al.* Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* 620, 434–444 (2023).
58. Mehlhoff, J. D. *et al.* Collateral fitness effects of mutations. *Proceedings of the National Academy of Sciences* 117, 11597–11607 (2020).
59. Ma, X. *et al.* Analysis of error profiles in deep next-generation sequencing data. *Genome Biology* 20, 1–15 (2019).
60. Livesey, B. J. & Marsh, J. A. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Molecular Systems Biology* 16, e9380 (2020).

61. Li, X. & Lehner, B. Biophysical ambiguities prevent accurate genetic prediction. *Nature Communications* 11, 4923 (2020).
62. Touw, W. G. *et al.* A series of PDB-related databanks for everyday needs. *Nucleic Acids Research* 43, D364–D368 (2015).
63. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules* 22, 2577–2637 (1983).
64. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021).
65. Carpenter, B. *et al.* Stan: A probabilistic programming language. *Journal of Statistical Software* 76 (2017).
66. Perkins, S. J. Protein volumes and hydration effects: the calculations of partial specific volumes, neutron scattering matchpoints and 280-nm absorption coefficients for proteins and glycoproteins from amino acid sequences. *European Journal of Biochemistry* 157, 169–180 (1986).
67. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* 157, 105–132 (1982).