OXFORD GENETICS

GENETICS, 2026, 233(2), iyag104
https://doi.org/10.1093/genetics/iyag104
Advance Access Publication Date: 21 April 2026
Investigation

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]

[1]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, United States
[2]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA 02115, United States
[3]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, United States
[4]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, United States
[5]Department of Physics, Harvard University, Cambridge, MA 02138, United States

*Corresponding author: Department of Systems Biology, Harvard Medical School, Boston, MA 02115, United States. Email: berk_alpay@g.harvard.edu.

Deep mutational scans have measured the effects of many mutations on a large variety of proteins. We use a collection of these scans to perform a meta-analysis of the effects of single amino acid substitutions, modeling the relative deleteriousness of each substitution with respect to the surface accessibility of the wildtype residue and the identities of the amino acids being exchanged. This simple model quantifies physicochemical trends underlying mutational effects across proteins and phenotypes, and manages to explain much of the observed variance.

Keywords: mutational effects; proteins; deep mutational scanning; Bayesian regression

Introduction

It is of fundamental interest in medicine (Stefl et al. 2013), molecular evolution (Maynard Smith 1970), and protein engineering (Bloom and Arnold 2009) to understand the effects of mutations on proteins. Accordingly, experimental techniques, most prominently deep mutational scanning (Fowler and Fields 2014), have been developed to measure the effects of mutations at high throughput on proteins of interest. Effects that can be assayed in this way include those on expression (Vanella et al. 2024), binding affinity (Starr et al. 2020), and catalytic activity (Jacquier et al. 2013; Vanella et al. 2024). Databases such as MaveDB (Esposito et al. 2019) and ProteinGym (Notin et al. 2023) have standardized and compiled the results of many deep mutational scans.

In parallel, machine learning models have been developed to predict the deleteriousness of mutations by comparing the mutated sequence to repositories of natural sequences (The UniProt 2017). They have traditionally used only sequence homologs (Ng and Henikoff 2001; Hopf et al. 2017; Laine et al. 2019), while protein language models (Meier et al. 2021; Su et al. 2023), recent alternatives, instead learn from a general corpus of sequences before tailoring predictions to the protein of interest. Predictions of machine learning models are very often compared against deep mutational scans for validation (Livesey and Marsh 2020).

While proteins are diverse (Berman et al. 2000) and their properties biophysically complex (Whisstock and Lesk 2003), trends in mutational effects have been noted. Early work with individual proteins revealed structural considerations in the tolerance of substitutions (Bowie et al. 1990). It is well-recognized for example that polar substitutions to buried residues tend to be disruptive (Perutz et al. 1965; Reidhaar-Olson and Sauer 1988), more so than substitutions to solvent-accessible residues (Schwehm et al. 1998; Guo et al. 2004). Similarly, substitutions to loops and turns are typically better tolerated than those to alpha helices and beta sheets (Guo et al. 2004), with proline-substitution posing a particular threat to these structured regions (Chou and Fasman 1974). Indeed, methods of predicting disease variants have often incorporated physicochemical features (Chasman and Adams 2001; Sunyaev et al. 2001; Adzhubei et al. 2010; Ponzoni et al. 2020), typically in addition to other data including sequence alignments.

The fairly recent profusion of deep mutational scanning data enables more precise and better-powered investigation of these physicochemical trends, an opportunity others have taken to deliver insightful analyses (Stein et al. 2019; Cagiada et al. 2021). With this data, Gray et al., for example, argued histidine and asparagine substitutions are the ones most demonstrative of the effects of other substitutions (Gray et al. 2017) as opposed to alanine scanning (Cunningham and Wells 1989). Dunham and Beltrao (2021) categorized sites in several dozens of proteins by patterns of mutational effects and interpreted those categories with respect to local structural context. Mutational effects have also been modeled directly with empirical amino acid exchangeability matrices (Munro and Singh 2020; Høie et al. 2022; Yu et al. 2024), and with the surface accessibility of the substituted residue (Sruthi et al. 2020; Tsishyn et al. 2025). Schulze and Lindorff-Larsen (2024) combined these features to infer a separate exchangeability matrix for buried and exposed residues in analyzing effects on proteins' 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

Received on 24 February 2026; accepted on 15 April 2026

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.

Materials and methods

We use the deep mutational scans (as well as study metadata, effect predictions by machine learning models, and AlphaFold2-predicted (Jumper et al. 2021) protein structures) provided by ProteinGym (Notin et al. 2023) version 1.2, which is composed of 217 assays (Supplementary Fig. 1) and nearly 700,000 single amino acid substitutions. Substitutions at each site are well-sampled; at most sites, the effects of substituting the wildtype with each of the 19 other proteinogenic amino acids have been measured (Supplementary Fig. 2a). Some amino acids, such as leucine, glutamic acid, and glycine, are 3- to 4-fold more commonly the wildtype than others such as tryptophan, cysteine, and histidine (Supplementary Fig. 2b).

Although effect scores are monotone in deleteriousness, due to methodological differences, between scans they can widely vary in scale (Dunham and Beltrao 2021; Schulze et al. 2026). Previous studies have accounted for this problem by restricting scans to those that use a specific experimental technique (Schulze and Lindorff-Larsen 2024), transforming scores toward a standard scale (Dunham and Beltrao 2021), and normalizing the ranks of scores within scans (Munro and Singh 2020). To analyze all scans together without transformation and to take into account potential differences in the kinds of sites represented between scans, 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 (Luce 1959; Plackett 1975); each substitution has a corresponding deleteriousness score which we infer, and 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 0 to 1, the latter representing maximal deleteriousness. (Although beneficial effects can be detected by deep mutational scans, we assume that the best effect is no effect.) 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.

The first model determines the score of a particular substitution purely with respect to the identities of the amino acids being exchanged (Fig. 1a). In this model, there are therefore a set of $20 \times 19 = 380$ deleteriousness scores, one for each possible exchange of amino acids. Specifically, we assume a Plackett–Luce model in which the score of a w to m mutation is logistic($\beta_{w,m}$), where independently $\beta_{w,m} \sim \text{Normal}(0, 1)$. We perform maximum a posteriori estimation of these scores in Stan (Carpenter et al. 2017). Each score lies between 0 and 1 and represents the aggregate deleteriousness of substituting one amino acid for another.

The second model, in addition to amino acid exchangeability, also includes surface accessibility, extracted with DSSP (Kabsch and Sander 1983; Touw et al. 2015) from the AlphaFold2-predicted structure. The deleteriousness score of each substitution still lies 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. Formally, the deleteriousness score of a w to m mutation whose wildtype residue has relative accessible surface area s is logistic($\beta_{w_i,m_i} + \beta_{\text{SA}} s_i$), where independently $\beta_{w,m}, \beta_{\text{SA}} \sim \text{Normal}(0, 1)$.

Results

Altogether, amino acid exchangeabilities explain a considerable proportion of the variance in the experimentally ranked effects (Fig. 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 (Fig. 2a). Cysteine's obstinacy as a wildtype residue 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 (Fig. 2b) (Munro and Singh 2020). 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, whereas proline-wildtype sites have already accepted this biophysical limitation. Similarly, a mutant cysteine residue would not be expected to form a disulfide bridge. More broadly, uncharged-to-charged substitutions tend to be more deleterious than charged-to-uncharged substitutions, possibly reflecting that uncharged amino acids, often buried in the hydrophobic core, are particularly intolerant of hydrophilic substitutes.

ProteinGym categorizes scans by the phenotype assessed: catalytic or biochemical activity, binding affinity to a target, expression, organismal growth rate, and stability. (Most stability scans are from a single study of many domains; Tsuboyama et al. 2023.) 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 (Supplementary Fig. 3). Notably, replacing hydrophilic amino acids with hydrophobic ones tends to be more deleterious than usual among binding assays.

The empirical trends of deleteriousness with respect to surface accessibility (Supplementary Fig. 4) indicate that this relationship is fairly consistent across substitutions, so we modeled it with a single parameter β_{SA}. The inferences of the exchangeability and surface accessibility model (Fig. 3a) align reasonably well with the empirical trends in surface accessibility. The inferred exchangeabilities are very similar, with Spearman correlation 0.91, to those of the exchangeability-only model. Figure 3b demonstrates the model's predictions.

The Spearman correlation of this model with experimental measurements varies widely across deep mutational scans but is more consistent among larger scans, where it is between approximately 0.3 and 0.6. The model is similar in average per-scan Spearman correlation across types of scans (around 0.4), but it is the highest (0.57) with respect to stability (Supplementary Fig. 5). Of the models we compare in Fig. 1b, the exchangeability and surface accessibility model bears the highest Spearman correlation

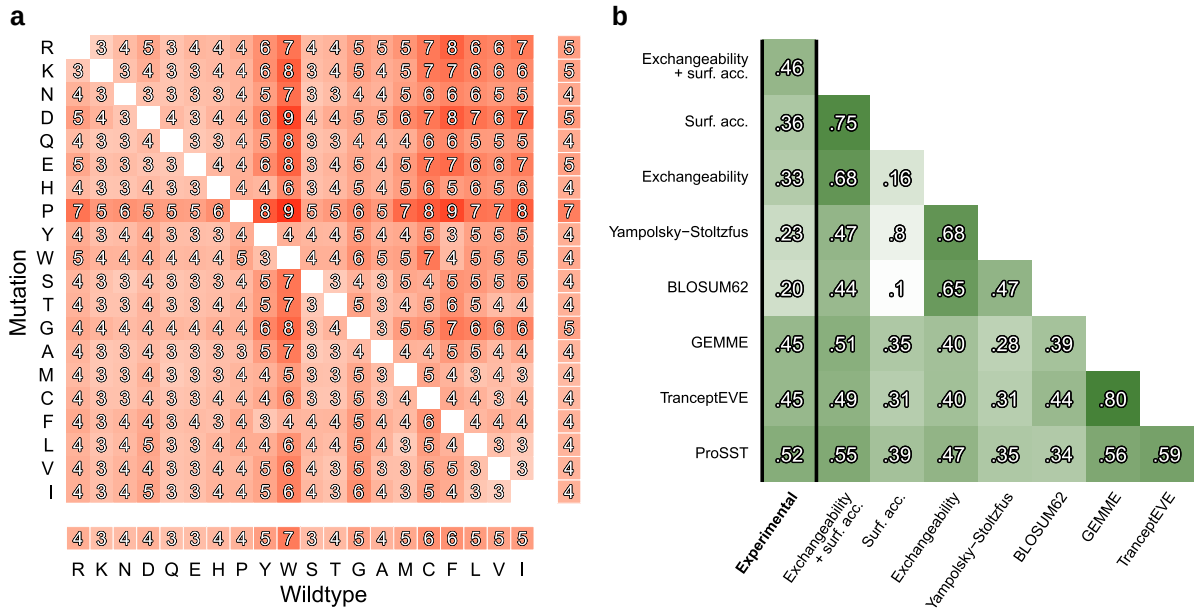

Fig. 1. a) Amino acid exchangeabilities as inferred from ranks of mutational effects within deep mutational scans. Marginal scores represent averages for the wildtype or mutant residues. Higher scores indicate greater deleteriousness, with scores scaled ×10 and rounded to the nearest integer. Amino acids are ordered according to Kyte and Doolittle's hydropathy index (Kyte and Doolittle 1982). b) The mean 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 10-fold cross-validation among proteins. Included are the Yampolsky–Stoltzfus exchangeability matrix (Yampolsky and Stoltzfus 2005) and BLOSUM62 (Henikoff and Henikoff 1992), and the sequence-based machine learning models are GEMME (Laine et al. 2019), an alignment-based model, TranceptEVE (Notin et al. 2022), an alignment-based and protein language model, and ProSST (Li et al. 2024), a structure and protein language model (Notin et al. 2023).

with predicted ΔΔG among a subset of 25 scans (Fig. 4a). Substitutions of glycine, surface residues, and residues in nonregular secondary structures appear to be particularly prone to prediction error, though the range in mean absolute rank error with respect to amino acid exchange, surface accessibility, and secondary structure is between about only 20–30% (Supplementary Fig. 6).

Surface accessibility alone explains the measured ranks better than exchangeability alone, and combining them considerably increases the variance explained to the level of sequence-based machine learning models (Fig. 1b). (We also computed the average per-scan Spearman correlation of MIF (Yang et al. 2023), an inverse folding model, with deep mutational 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. In addition to comparing whole score sets, we also computed how well the models recall the 10% least deleterious mutations measured in each scan (Fig. 4b). The median recall for all models is below 20%. The recall of the exchangeability and surface accessibility model approaches but remains below that of sequence-based models, indicating that though this is a difficult task, data of evolutionary conservation is particularly useful for identifying relatively permissive loci.

Discussion

Through statistical modeling of a large variety of deep mutational scans, we analyzed aggregate trends in the relative deleteriousness of single amino acid substitutions with respect to amino acid exchangeability and surface accessibility. We found that these features, on par with the predictions of sequence-based machine learning models with many more parameters, explain much of the variance in deleteriousness in deep mutational scans.

Models that combine site-wise evolutionary conservation with amino acid exchangeability (Munro and Singh 2020), and site-wise conservation with solvent accessibility (Tsishyn et al. 2025) are also relatively simple and explain a good deal of mutational effects across a variety 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 strong prior over mutational effects which we believe may be useful toward various ends, perhaps as a prior for parameters of machine learning models, especially when homologous sequence data is limited, and as a baseline (Høie et al. 2022) for models that use more specific features.

Empirical departures from the prior could also indicate functionally significant residues (Schulze and Lindorff-Larsen 2024; Rao et al. 2026). For example, the residues of TP53 whose average deleterious ranks are the most underestimated (relative to Kotler et al. 2018) of positions 102 through 292 are, in increasing order: R280, A276, S241, R248, and G244, of which R248, and R280 are critical DNA-binding contacts (Cho et al. 1994). The most underestimated residues of the SARS-CoV-2 spike protein (relative to Starr et al. 2020) of residues 331 through 531 are F486, Y505, N343, G502, and T500, of which all are ACE2 contacts (Lan et al. 2020) except N343, which is a glycosylation site involved in ACE2 binding (Sztain et al. 2021).

We compared our models to sequence-based models as a measure of how well they could explain the relative deleteriousness of single substitutions. Our models cannot, however, substitute these machine learning methods, for several reasons. Sequence-based models are unsupervised, can perform other tasks (such as structure prediction; Marks et al. 2011 and sequence generation; Riesselman et al. 2018; Su et al. 2023), and can capture epistatic effects (Hopf et al. 2017; Laine et al. 2019)

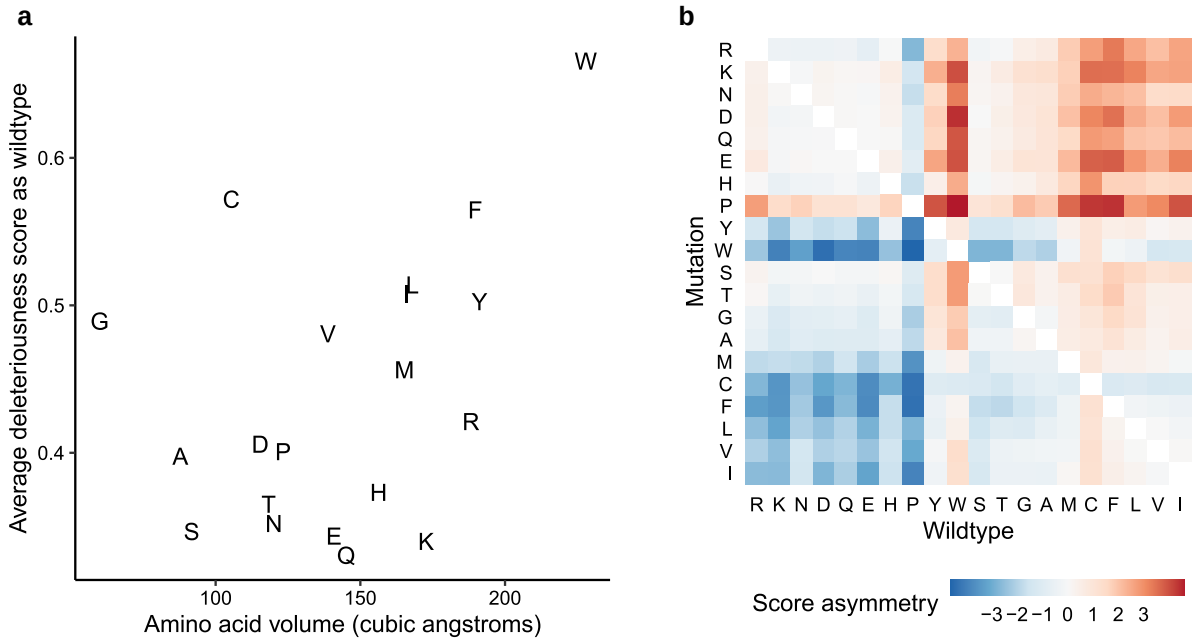

Fig. 2. a) The average exchangeability of each amino acid wildtype versus its volume as calculated by Perkins (1986). b) For each amino acid substitution *a* for *b*, the deleteriousness score relative to the reverse substitution *b* for *a*. The greater the score asymmetry, the greater the deleteriousness of the forward substitution than the reverse.

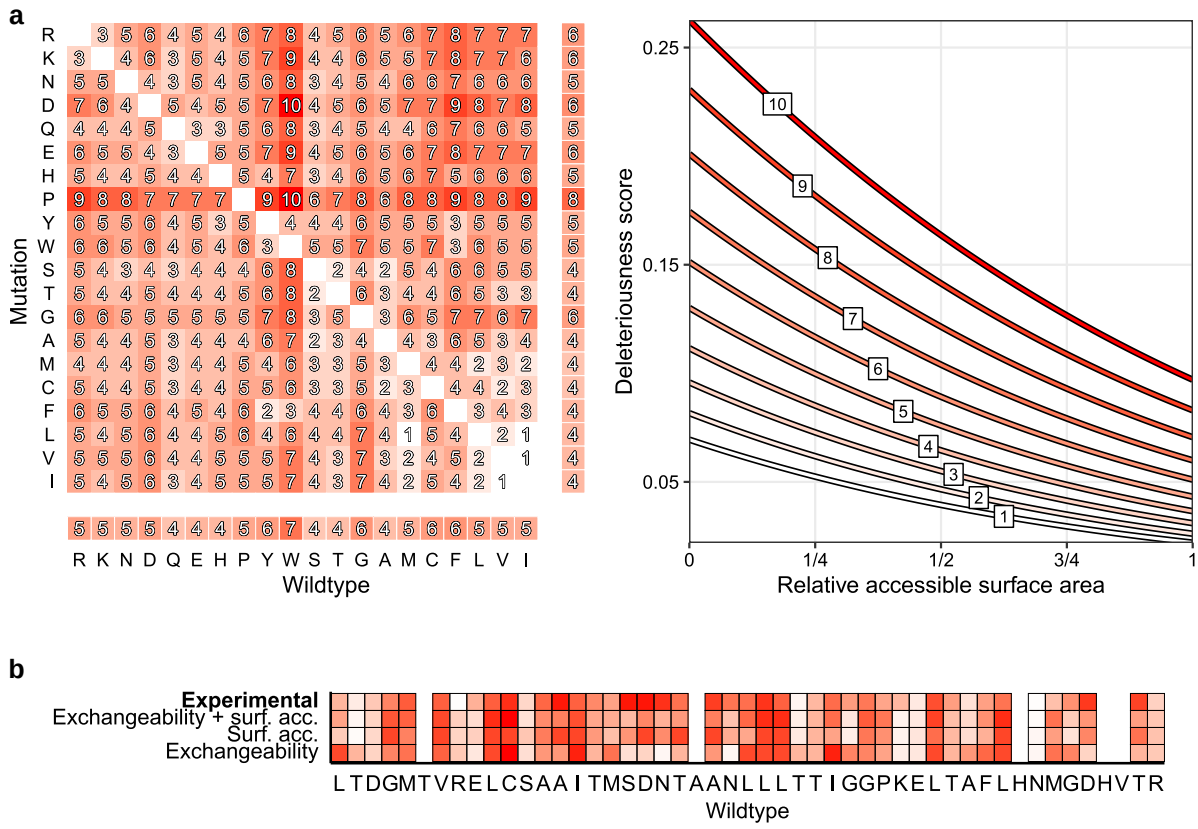

Fig. 3. a) Deleteriousness scores with respect to the amino acids being exchanged and the wildtype residue's surface accessibility. For illustration, amino acid scores are binned from one through ten and the deleteriousness curve with respect to surface accessibility is shown using the central score 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. (2014). (Some loci lack experimental measurements.) The Spearman correlation of the full model with the experiment is 0.68 in this region.

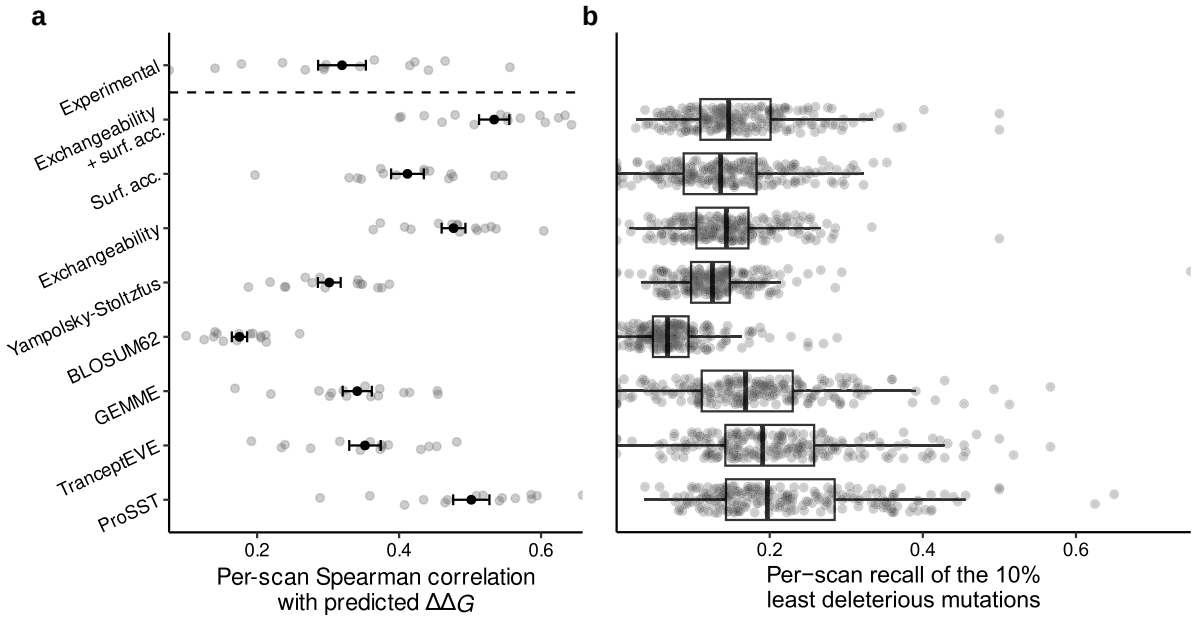

Fig. 4. a) The correlation of models with Rosetta-predicted ΔΔG (Leman et al. 2020) for 25 scans across 15 matching proteins in ProteinGym (Supplementary Table 1) as calculated by Høie et al. (2022). Spearman correlations of scans of the same protein are averaged before the grand mean and standard error are calculated. b) Per deep mutational scan, the fraction of the 10% least deleterious mutations recalled as such by each model.

common in proteins (Starr and Thornton 2016; Buda et al. 2023). It is also worth noting that the inferences of our models are biased toward those proteins and mutations that have been sampled by deep mutational scans and that sequence-based models can in principle benefit from a wider variety of training data.

An immediate product of our statistical analysis was an empirical amino acid exchangeability matrix that quantifies the relative deleteriousness of each of the 380 possible amino acid substitutions. The matrix, like other empirical exchangeability matrices (Munro and Singh 2020), 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, and contrasts with substitution matrices based on evolutionary conservation, eg BLOSUM (Henikoff and Henikoff 1992) and PAM (Dayhoff et al. 1978), and physicochemical distance matrices (Grantham 1974; Miyata et al. 1979; Kawashima et al. 2007), which are symmetric and do not consider mutational effects directly. Substitution matrices reflect factors like codon bias and the structure of the genetic code (Yampolsky and Stoltzfus 2005) and are not intended for mutation effect prediction (Ng and Henikoff 2001), 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 one dozen systematic exchange studies (Yampolsky and Stoltzfus 2005), precursors to deep mutational scans. 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 (Fig. 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 (2020) and one by Høie et al. (2022).

Our models explain much of the relative deleteriousness of mutations without accounting for the assayed phenotype, and the inferences from training on assay types separately are similar to those learned from the whole data. These results are consistent with there being general properties of proteins, including stability (Bloom and Arnold 2009; Stein et al. 2019; Gerasimavicius et al. 2023) (but also perhaps aggregation and misfolding potential Mehlhoff et al. 2020; Cagiada et al. 2021), 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. Indeed, the predictions of our model track particularly well with predicted ΔΔG, and the stability scans are the ones best-explained by exchangeability and surface accessibility.

We have shown that amino acid exchangeability and a surface accessibility parameter explain much of the effects of single substitutions across various deep mutational scans. Other physicochemical features could be incorporated, though one should expect natural limits on correspondence with experimental scores, which can be noisy (Ma et al. 2019; Livesey and Marsh 2020) and vary between replicates. Idiosyncratic interactions between sites and the intricacies of protein function (Li and Lehner 2020; Buda et al. 2023), as well as such external considerations as molecular interactions, add to the complexity of intuiting the effects of mutations using small models. 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.

Data availability

Code for this article, as well as inferred model parameters, can be accessed at https://github.com/berkalpay/metamutation. ProteinGym version 1.2 is archived at https://doi.org/10.5281/zenodo.14997691.

Supplemental material available at GENETICS online.

Acknowledgments

Thank you to Pascal Notin for useful discussions about ProteinGym and to Wendy Valencia-Montoya for her feedback on the text.

Funding

B.A.A. was supported by an NSF Graduate Research Fellowship under Grant No. DGE-2140743.

Conflicts of interest

None declared.

Literature cited

Adzhubei IA *et al.* 2010. A method and server for predicting damaging missense mutations. Nat Methods. 7:248–249. https://doi.org/10.1038/nmeth0410-248.

Berman HM *et al.* 2000. The protein data bank. Nucleic Acids Res. 28:235–242. https://doi.org/10.1093/nar/28.1.235.

Bloom JD, Arnold FH. 2009. In the light of directed evolution: pathways of adaptive protein evolution. Proc Natl Acad Sci U S A. 106:9995–10000. https://doi.org/10.1073/pnas.0901522106.

Bowie JU, Reidhaar-Olson JF, Lim WA, Sauer RT. 1990. Deciphering the message in protein sequences: tolerance to amino acid substitutions. Science. 247:1306–1310. https://doi.org/10.1126/science.2315699.

Buda K, Miton CM, Tokuriki N. 2023. Pervasive epistasis exposes intramolecular networks in adaptive enzyme evolution. Nat Commun. 14:8508. https://doi.org/10.1038/s41467-023-44333-5.

Cagiada M *et al.* 2021. Understanding the origins of loss of protein function by analyzing the effects of thousands of variants on activity and abundance. Mol Biol Evol. 38:3235–3246. https://doi.org/10.1093/molbev/msab095.

Carpenter B *et al.* 2017. Stan: a probabilistic programming language. J Stat Softw. 76:1–32. https://doi.org/10.18637/jss.v076.i01.

Chasman D, Adams RM. 2001. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol. 307:683–706. https://doi.org/10.1006/jmbi.2001.4510.

Cho Y, Gorina S, Jeffrey PD, Pavletich NP. 1994. Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. Science. 265:346–355. https://doi.org/10.1126/science.8023157.

Chou PY, Fasman GD. 1974. Prediction of protein conformation. Biochemistry. 13:222–245. https://doi.org/10.1021/bi00699a002.

Cunningham BC, Wells JA. 1989. High-resolution epitope mapping of HGH-receptor interactions by alanine-scanning mutagenesis. Science. 244:1081–1085. https://doi.org/10.1126/science.2471267.

Dayhoff M, Schwartz R, Orcutt B. 1978. A model of evolutionary change in proteins. Atlas Pro Seq Structure. 5:345–352.

Dunham AS, Beltrao P. 2021. Exploring amino acid functions in a deep mutational landscape. Mol Syst Biol. 17:e10305. https://doi.org/10.15252/msb.202110305.

Esposito D *et al.* 2019. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. Genome Biol. 20:1–11. https://doi.org/10.1186/s13059-018-1612-0.

Firnberg E, Labonte JW, Gray JJ, Ostermeier M. 2014. A comprehensive, high-resolution map of a gene's fitness landscape. Mol Biol Evol. 31:1581–1592. https://doi.org/10.1093/molbev/msu081.

Fowler DM, Fields S. 2014. Deep mutational scanning: a new style of protein science. Nat Methods. 11:801–807. https://doi.org/10.1038/nmeth.3027.

Gerasimavicius L, Livesey BJ, Marsh JA. 2023. Correspondence between functional scores from deep mutational scans and predicted effects on protein stability. Protein Sci. 32:e4688. https://doi.org/10.1002/pro.4688.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science. 185:862–864. https://doi.org/10.1126/science.185.4154.862.

Gray VE, Hause RJ, Fowler DM. 2017. Analysis of large-scale mutagenesis data to assess the impact of single amino acid substitutions. Genetics. 207:53–61. https://doi.org/10.1534/genetics.117.300064.

Guo HH, Choe J, Loeb LA. 2004. Protein tolerance to random amino acid change. Proc Natl Acad Sci U S A. 101:9205–9210. https://doi.org/10.1073/pnas.0403255101.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 89:10915–10919. https://doi.org/10.1073/pnas.89.22.10915.

Høie MH, Cagiada M, Frederiksen AHB, Stein A, Lindorff-Larsen K. 2022. Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. Cell Rep. 38:110207. https://doi.org/10.1016/j.celrep.2021.110207.

Hopf TA *et al.* 2017. Mutation effects predicted from sequence covariation. Nat Biotechnol. 35:128–135. https://doi.org/10.1038/nbt.3769.

Jacquier H *et al.* 2013. Capturing the mutational landscape of the beta-lactamase TEM-1. Proc Natl Acad Sci U S A. 110:13067–-13072. https://doi.org/10.1073/pnas.1215206110.

Jumper J *et al.* 2021. Highly accurate protein structure prediction with AlphaFold. Nature. 596:583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 22:2577–2637. https://doi.org/10.1002/bip.360221211.

Kawashima S *et al.* 2007. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 36:D202–D205. https://doi.org/10.1093/nar/gkm998.

Kotler E *et al.* 2018. A systematic p53 mutation library links differential functional impact to cancer mutation pattern and evolutionary conservation. Mol Cell. 71:178–190. https://doi.org/10.1016/j.molcel.2018.06.012.

Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 157:105–132. https://doi.org/10.1016/0022-2836(82)90515-0.

Laine E, Karami Y, Carbone A. 2019. GEMME: a simple and fast global epistatic model predicting mutational effects. Mol Biol Evol. 36:2604–2619. https://doi.org/10.1093/molbev/msz179.

Lan J *et al.* 2020. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature. 581:215–220. https://doi.org/10.1038/s41586-020-2180-5.

Leman JK *et al.* 2020. Macromolecular modeling and design in Rosetta: recent methods and frameworks. Nat Methods. 17:665–680. https://doi.org/10.1038/s41592-020-0848-2.

Li M *et al.* 2024. ProSST: protein language modeling with quantized structure and disentangled attention. Adv Neural Inf Process Syst. 37:35700–35726. https://doi.org/10.52202/079017-1126.

Li X, Lehner B. 2020. Biophysical ambiguities prevent accurate genetic prediction. Nat Commun. 11:4923. https://doi.org/10.1038/s41467-020-18694-0.

Livesey BJ, Marsh JA. 2020. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. Mol Syst Biol. 16:e9380. https://doi.org/10.15252/msb.20199380.

Luce RD. 1959. Individual choice behavior: a theoretical analysis. Wiley.

Ma X *et al.* 2019. Analysis of error profiles in deep next-generation sequencing data. Genome Biol. 20:1–15. https://doi.org/10.1186/s13059-018-1612-0.

Marks DS *et al.* 2011. Protein 3D structure computed from evolutionary sequence variation. PLoS One. 6:e28766. https://doi.org/10.1371/journal.pone.0028766.

Maynard Smith J. 1970. Natural selection and the concept of a protein space. Nature. 225:563–564. https://doi.org/10.1038/225563a0.

Mehlhoff JD *et al.* 2020. Collateral fitness effects of mutations. Proc Natl Acad Sci U S A. 117:11597–11607. https://doi.org/10.1073/pnas.1918680117.

Meier J *et al.* 2021. Language models enable zero-shot prediction of the effects of mutations on protein function. Adv Neural Inf Process Syst. 34:29287–29303. https://doi.org/10.1101/2021.07.09.450648.

Miyata T, Miyazawa S, Yasunaga T. 1979. Two types of amino acid substitutions in protein evolution. J Mol Evol. 12:219–236. https://doi.org/10.1007/BF01732340.

Munro D, Singh M. 2020. DeMaSk: a deep mutational scanning substitution matrix and its use for variant impact prediction. Bioinformatics. 36:5322–5329. https://doi.org/10.1093/bioinformatics/btaa1030.

Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. Genome Res. 11:863–874. https://doi.org/10.1101/gr.176601.

Notin P *et al.* 2022. TranceptEVE: combining family-specific and family-agnostic models of protein sequences for improved fitness prediction [preprint]. bioRxiv. https://doi.org/10.1101/2022.12.07.519495.

Notin P *et al.* 2023. ProteinGym: large-scale benchmarks for protein fitness prediction and design. Adv Neural Inf Process Syst. 36:64331–64379. https://doi.org/10.1101/2023.12.07.570727.

Perkins SJ. 1986. 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. Eur J Biochem. 157:169–180. https://doi.org/10.1111/j.1432-1033.1986.tb09653.x.

Perutz M, Kendrew J, Watson H. 1965. Structure and function of haemoglobin: II. Some relations between polypeptide chain configuration and amino acid sequence. J Mol Biol. 13:669–678. https://doi.org/10.1016/S0022-2836(65)80134-6.

Plackett RL. 1975. The analysis of permutations. J R Stat Soc Ser C Appl Stat. 24:193–202. https://doi.org/10.2307/2346567.

Ponzoni L, Peñaherrera DA, Oltvai ZN, Bahar I. 2020. Rhapsody: predicting the pathogenicity of human missense variants. Bioinformatics. 36:3084–3092. https://doi.org/10.1093/bioinformatics/btaa127.

Rao J *et al.* 2026. Rosace-AA: enhancing interpretation of deep mutational scanning data with amino acid substitution and position--specific insights. Bioinform Adv. 6:vbaf218. https://doi.org/10.1093/bioadv/vbaf218.

Reidhaar-Olson JF, Sauer RT. 1988. Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. Science. 241:53–57. https://doi.org/10.1126/science.3388019.

Riesselman AJ, Ingraham JB, Marks DS. 2018. Deep generative models of genetic variation capture the effects of mutations. Nat Methods. 15:816–822. https://doi.org/10.1038/s41592-018-0138-4.

Schulze TK, Blaabjerg LM, Cagiada M, Lindorff-Larsen K. 2026. Supervised learning of protein variant effects across large-scale mutagenesis datasets. Protein Sci. 35:e70526. https://doi.org/10.1002/pro.70526.

Schulze TK, Lindorff-Larsen K. 2024. Effects of residue substitutions on the cellular abundance of proteins. eLife. 13:RP103721. https://doi.org/10.7554/eLife.103721.2.

Schwehm JM, Kristyanne ES, Biggers CC, Stites WE. 1998. Stability effects of increasing the hydrophobicity of solvent-exposed side chains in staphylococcal nuclease. Biochemistry. 37:6939–6948. https://doi.org/10.1021/bi9725069.

Sruthi CK, Balaram H, Prakash MK. 2020. Toward developing intuitive rules for protein variant effect prediction using deep mutational scanning data. ACS Omega. 5:29667–29677. https://doi.org/10.1021/acsomega.0c02402.

Starr TN *et al.* 2020. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell. 182:1295–1310. https://doi.org/10.1016/j.cell.2020.08.012.

Starr TN, Thornton JW. 2016. Epistasis in protein evolution. Protein Sci. 25:1204–1218. https://doi.org/10.1002/pro.2897.

Stefl S, Nishi H, Petukh M, Panchenko AR, Alexov E. 2013. Molecular mechanisms of disease-causing missense mutations. J Mol Biol. 425:3919–3936. https://doi.org/10.1016/j.jmb.2013.07.014.

Stein A, Fowler DM, Hartmann-Petersen R, Lindorff-Larsen K. 2019. Biophysical and mechanistic models for disease-causing protein variants. Trends Biochem Sci. 44:575–588. https://doi.org/10.1016/j.tibs.2019.01.003.

Su J *et al.* 2023. SaProt: protein language modeling with structure-aware vocabulary [preprint]. bioRxiv. https://doi.org/10.1101/2023.10.01.560349.

Sunyaev S *et al.* 2001. Prediction of deleterious human alleles. Hum Mol Genet. 10:591–597. https://doi.org/10.1093/hmg/10.6.591.

Sztain T *et al.* 2021. A glycan gate controls opening of the SARS-CoV-2 spike protein. Nat Chem. 13:963–968. https://doi.org/10.1038/s41557-021-00758-3.

The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45:D158–D169. https://doi.org/10.1093/nar/gkw1099.

Touw WG *et al.* 2015. A series of PDB-related databanks for everyday needs. Nucleic Acids Res. 43:D364–D368. https://doi.org/10.1093/nar/gku1028.

Tsishyn M, Hermans P, Pucci F, Rooman M. 2025. Residue conservation and solvent accessibility are (almost) all you need for predicting mutational effects in proteins. Bioinformatics. 41:btaf322. https://doi.org/10.1093/bioinformatics/btaf322.

Tsuboyama K *et al.* 2023. Mega-scale experimental analysis of protein folding stability in biology and design. Nature. 620:434–444. https://doi.org/10.1038/s41586-023-06328-6.

Vanella R *et al.* 2024. Understanding activity-stability tradeoffs in biocatalysts by enzyme proximity sequencing. Nat Commun. 15:1807. https://doi.org/10.1038/s41467-024-45630-3.

Whisstock JC, Lesk AM. 2003. Prediction of protein function from protein sequence and structure. Q Rev Biophys. 36:307–340. https://doi.org/10.1017/s0033583503003901.

Yampolsky LY, Stoltzfus A. 2005. The exchangeability of amino acids in proteins. Genetics. 170:1459–1472. https://doi.org/10.1534/genetics.104.039107.

Yang KK, Zanichelli N, Yeh H. 2023. Masked inverse folding with sequence transfer for protein representation learning. Protein Eng Des Sel. 36:gzad015. https://doi.org/10.1093/protein/gzad015.

Yu T *et al.* 2024. FUSE: improving the estimation and imputation of variant impacts in functional screening. Cell Genom. 4:100667. https://doi.org/10.1016/j.xgen.2024.100667.

Editor: A. Moses