Title

Uncovering differential tolerance to deletions versus substitutions with a protein language model

ISBN

9798186184294

Author

Goldman, Grant

Publication Date

2025-12-17

Peer reviewed|Thesis/dissertation

Uncovering differential tolerance to deletions versus substitutions with a protein language model

by
Grant Goldman

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Ryan Hernandez
497B1075667F496...

Ryan Hernandez
Chair

DocuSigned by:
Vasilis Ntranos
DocuSigned by:
FD...

Vasilis Ntranos

DocuSigned by:
Willow Coyote-Maestas
4B590B75F57548D...

Willow Coyote-Maestas

Committee Members

Acknowledgements

This work was partially funded by the National Institutes of Health, National Human Genome Research Institute (NIH/NHGRI; grant R01AI129731) and the National Institute of General Medical Sciences (NIH/NIGMS; grant T32GM150479).

Thank you to Vasilis Ntranos for his guidance and support throughout my PhD. Thank you to Prathamesh Chati for his insight and teamwork through this project. Thank you to my committee members, Ryan Hernandez, Willow Coyote-Maestas, for their helpful feedback. Thank you to the participants of the BMI Research in Progress Seminar, the 2024 Mutational Scanning Symposium, and the QBC Retreats, where I presented preliminary results of this study. Thank you to past and present members of the Ntranos Lab (Tuan Dinh, Joe Germino, Connor Tomaka, Anish Amara, Moonkyung Choi) and the UCSF Diabetes Center for their friendship and camaraderie. Thank you to Rebecca Dawson for her dedicated administrative support of the BMI program.

Dedication

This work is dedicated to my parents, grandparents, siblings, cousins, and friends, all of whom made me into the person I am today.

Abstract

Deep mutational scanning (DMS) experiments have been successfully leveraged to understand genotype to phenotype mapping. However, the overwhelming majority of DMS have focused on amino acid substitutions. Thus, it remains unclear how indels differentially shape the fitness landscape relative to substitutions. In order to further our understanding of the relationship between substitutions and deletions, we leveraged a protein language model to analyze every single amino acid deletion in the human proteome. We discovered hundreds of thousands of sites that display opposing behavior for deletions versus substitutions: sites that can tolerate being substituted but not deleted, or vice versa. We identified secondary structural elements and sequence context to be important mediators of differential tolerance. Our results underscore the value of deletion-substitution comparisons at the genome-wide scale, provide novel insights into how substitutions could systematically differ from deletions, and showcase the power of protein language models to generate biological hypotheses in-silico.

Table of Contents

Introduction

Understanding the effects of non-synonymous mutations is a critical challenge in fields such as human genetics and protein engineering.[1,2] To date, a growing body of computational and experimental literature has assessed the consequences of amino acid substitutions among a wide array of proteins,[3–6] yielding fundamental insights into how missense mutations are tolerated within a sequence context. However, other types of mutations, such as in-frame deletions, are far less understood.[7,8] Growing evidence suggests that deletions play a critical role in protein evolution and disease,[8,9] but it is unclear how a site's substitution tolerance may differ from its deletion tolerance, if at all.

Although notable progress has been made in expanding experimental frameworks to incorporate deletions,[7,10–15] they still remain vastly understudied relative to substitutions. According to a comprehensive survey of existing DMS data,[6] there are only a few thousand in-frame deletions, compared to over two million substitutions measured in the literature. Thus, current experimental data provides only a narrow window into the deletion fitness landscape. This discrepancy is due to limitations with existing DMS methodologies: inverse PCR for deletions is more resource and labor intensive than for substitutions, transposons have sequence biases that may result in an incomplete library, and oligo synthesis approaches are highly specific to the target protein and do not generalize well.[7]

Computational variant effect predictors (VEP) pose an attractive supplement to sparse experimental data on deletions. In particular, protein language models have recently emerged as efficient, scalable, and accurate predictors of mutation effects.[3,16–19] Here we propose using a

protein language model as a proxy of experimental deletion scans, allowing us to comprehensively assess their genome-wide effects. We argue that, while individual predictions may be noisy, collectively, at the genome-wide scale, they can generate a more representative view of the protein fitness landscape than limited experimental data alone. To this end, we employed ESM-1b[20] to score every possible single amino acid deletion across all 20,298 canonical human protein sequences in the UniProt database,[21] resulting in a dataset of over 11.2 million unique deletion effects (one for each site captured in the data).

To showcase the use of protein language models as *in-silico* hypothesis generation tools, we focus on the property of differential tolerance – whether a site can tolerate a substitution but not a deletion, or vice versa (**Fig. 1**). Using our ESM-generated data, we identified approximately 184,000 sites that meet this definition. We find that secondary structural elements and sequence context are strongly associated with differential tolerance. We further find evidence of differential tolerance in existing DMS studies, underscoring the importance of considering both deletions and substitutions in studies of protein fitness.

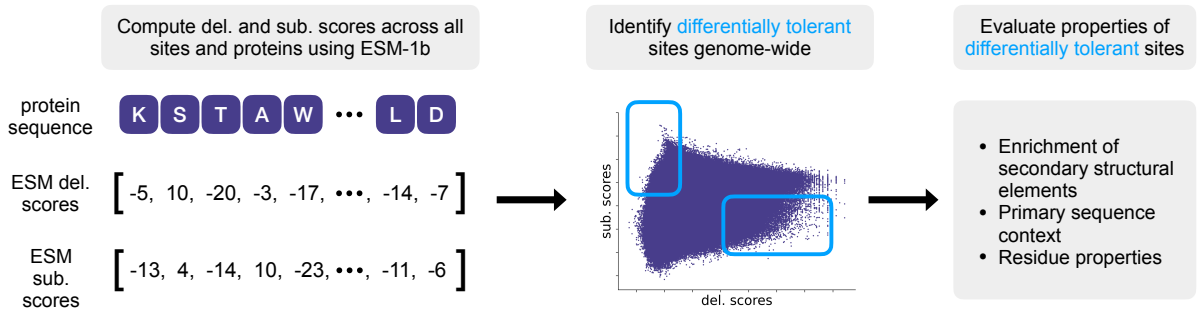


Figure 1: Genome-wide comparison of deletions versus substitutions with ESM. Overview of data generation process with ESM-1b, identification of sites with differential tolerance, and downstream analysis.

Results

Systematic comparison of substitutions versus deletions identifies sites with differential tolerance

We used our in-silico deletion dataset to estimate every site's simultaneous tolerance to deletions and substitutions. We score substitutions by comparing the probabilities of the mutant and wildtype residues in the context of the protein sequence. Deletions, alternatively, are scored by comparing the likelihood of the entire mutant sequence versus the entire wildtype sequence (**Methods**).

We have previously shown that ESM-1b has state-of-the-art performance for predicting clinical annotations for substitutions and deletions, in addition to experimental assays of substitutions.[3] Here, we additionally show ESM-1b's superior performance predicting experimental effects of deletions relative to other variant effect predictors, including the ESM2 (650 million parameter) model (**Supplementary Fig. S1**).

For an individual site, while deletions and substitutions are generally correlated, our deletion scan identified many notable exceptions (**Fig. 2a**). In particular, we partitioned the substitution and deletion scores into their respective top and bottom quartiles and found 184,000 sites that rank among the 25% most damaging deletions *and* the 25% most neutral substitutions (red quadrant, **Fig. 2a**), or vice versa (green quadrant, **Fig. 2a**). We define the red quadrant as sites that are deletion-intolerant and substitution-tolerant and refer to them as $D^{-}S^{+}$. Similarly, we

define the green quadrant as sites that are deletion-tolerant and substitution-intolerant and refer to them as D^+S^-.

Using ClinVar[22] annotations, we observe that every deletion score in the D^-S^+ quadrant is predicted more damaging than the median score among ClinVar pathogenic deletions, while substitution scores of these sites are predicted more neutral than the median score among ClinVar benign substitutions. Analogous results hold for the D^+S^- quadrant (**Fig. 2a**).

As an alternative to our quartile-based differential tolerance thresholds, we considered a continuous measure of differential tolerance (**Supplementary Fig. S2**). This approach yielded substantial overlap with our thresholding-based method. As such, we used the sites nominated via thresholding for all results presented here.

Notably, differential tolerance is observed in contexts beyond our genome-wide human data (**Fig. 2b**). In particular, we extracted sites from a large mutagenesis experiment of human, non-human, and de novo designed functional domains.[10] We then used ESM to predict deletion and substitution tolerance among the assayed sites. When compared against our genome-wide thresholds, we find the presence of both differential tolerance conditions among sites in these domains. This indicates that differential tolerance extends beyond the human proteome. Although the experimental domains tend to under-sample differentially tolerant sites compared to our genome-wide dataset, their presence nonetheless highlights the importance of differential tolerance when designing mutagenesis experiments.

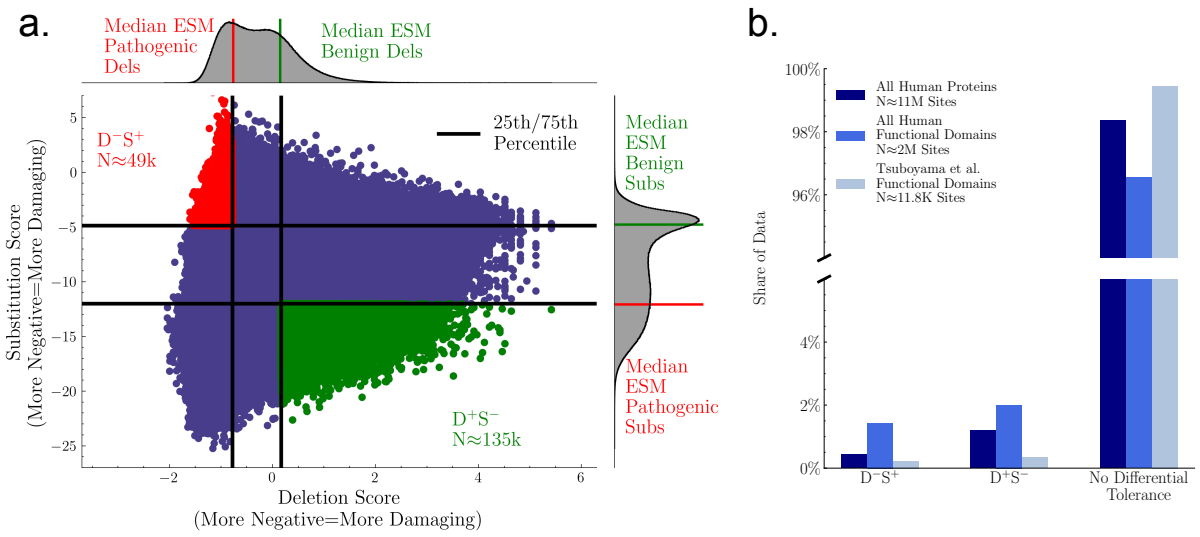


Figure 2: Identification of sites with differential tolerance. (a) Scatterplot (N= 11,291,148) of all sites' deletion versus substitution tolerance, sites with differential tolerance highlighted. Marginal distribution of genome-wide deletion scores (gray, top x-axis), with median ESM score of ClinVar benign deletions and pathogenic deletions highlighted. Analogous data for substitutions on the right y-axis. See **Methods** for computation of substitution and deletion scores. **(b)** Percent of sites with differential tolerance in ESM data (genome-wide and domains only) and experimental data (domains only, N=11,777).[10]

Secondary structural elements mediate $D^{-}S^{+}$ differential tolerance

To understand what drives differential tolerance, we compared the frequency of secondary structural elements within the $D^{-}S^{+}$ and $D^{+}S^{-}$ quadrants against their background, genome-wide frequencies.[21,23,24] We first focused on the $D^{-}S^{+}$ quadrant and observed an enrichment of sites contained within alpha-helices (34.8% of sites genome-wide versus 62.2% among $D^{-}S^{+}$ sites, **Fig. 3a** and **Supplementary Fig. S3**). We reasoned this is a consequence of the periodic nature of the structure, which completes a full turn every 3.6 residues; substitutions conserve the number of residues in the turn, but deletions do not.

Indeed, past work has investigated the effects of deletions within alpha-helices, noting that they can lead to destabilization or other functional alterations.[10,11,13,25–30] However, our analysis quantifies differential tolerance between substitutions and deletions within the structure at the

genome-wide scale. Moreover, jointly considering deletions and substitutions clarifies this effect beyond considering them independently (**Supplementary Fig. S4**).

We also examined the model's internal consistency with respect to $D^{-}S^{+}$ predictions by analyzing all non-synonymous substitutions (20*19=380 unique possible pairs) occurring in helical versus non-helical regions. For each unique amino acid substitution, we used ESM to compute the average effect of the mutation inside versus outside helices. We found that size-conservative substitutions produce more similar effects regardless of structural context, while non-size-conservative substitutions tend to be more damaging within helices.[31] These findings suggest that helices are particularly sensitive to bulk-altering mutations, which helps explain their enrichment among $D^{-}S^{+}$ sites (**Fig. 3b, Methods**). It is important to note, however, that this relationship is observed most clearly when restricting to high-pLDDT helices (pLDDT > 0.9, **Fig. 3b**). The effect is more ambiguous when considering all helices (**Supplementary Fig. S5**). We reason that, because $D^{-}S^{+}$ sites are enriched in high-pLDDT regions (**Supplementary Fig. S6**), the pLDDT restriction identifies the subset of helices most sensitive to bulk-altering variants.

To probe the sensitivity of helices to length-altering mutations, we compared the ESM-predicted effects of insertions versus substitutions across 1,632 structurally diverse proteins (**Methods**). Relative to the effects of substitutions, insertions tend to be significantly more damaging within helices than outside helices ($p<0.001$, **Fig. 3c**). Similarly, while insertions, deletions, and substitutions all tend to be more damaging within helices, this property is far more pronounced for insertions and deletions, which are length-altering, compared to substitutions (**Supplementary Fig. S7**).

As a final validation, we analyzed experimental stability measurements[10] and found that insertions and deletions within helices are approximately three times as destabilizing as substitutions within helices. Helices show the clearest signal of differential tolerance to insertions and deletions (relative to substitutions) among all structural elements considered (**Fig. 3d**).

Lastly, we considered other structural features (contact number, AlphaFold pLDDT, rSASA) with respect to differential tolerance (**Supplementary Fig. S6**). Contact number does not show a strong correlation with differential tolerance. Both differential tolerance conditions have higher pLDDT values, on average, than the genome-wide background. While rSASA appears to separate more distinctly by differential tolerance condition, we note a substantial similarity between rSASA distributions by differential tolerance condition and rSASA by secondary structure.

Taken together, these results clarify the tendency of alpha-helices to be overrepresented among sites that are simultaneously deletion-intolerant and substitution-tolerant. Our analyses indicate that ESM-1b has recapitulated a critical underlying property of alpha-helices, namely that their periodicity renders size-altering mutations more deleterious compared to size-conservative mutations.

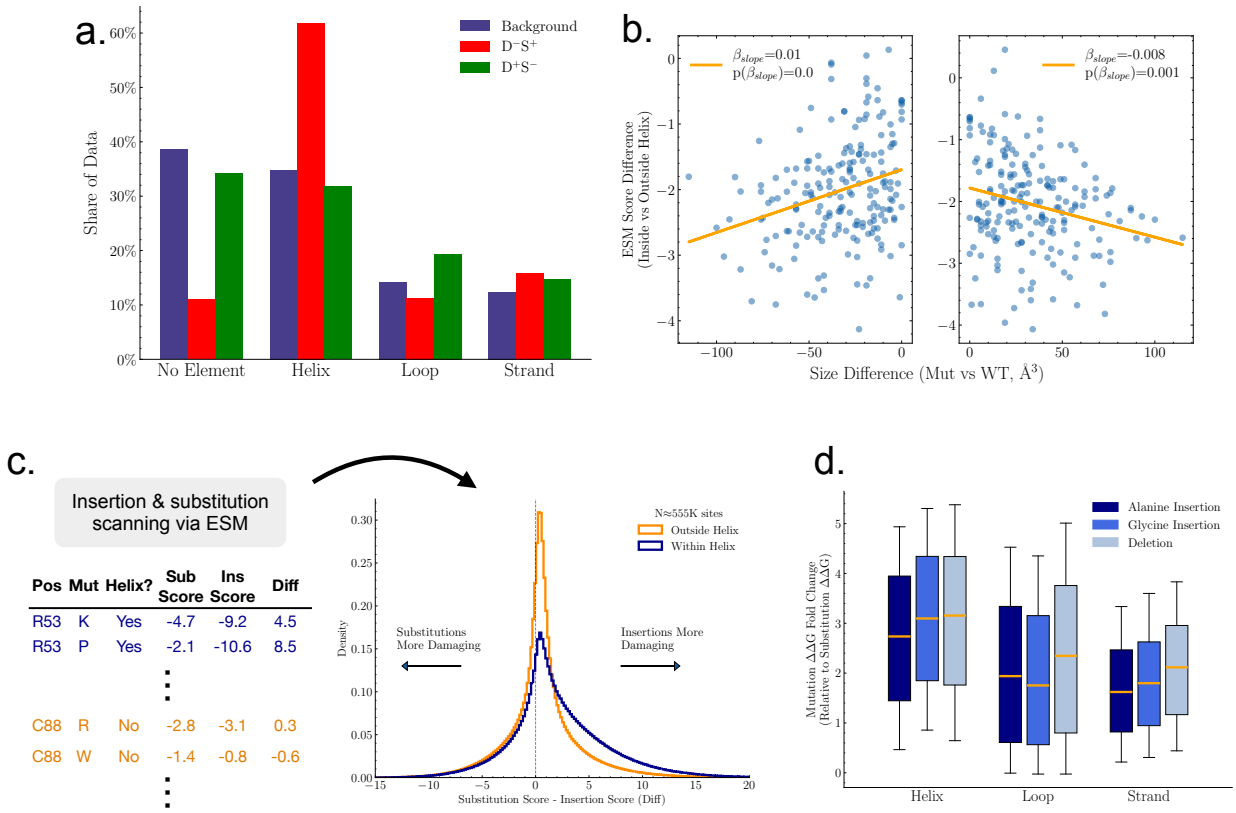


Figure 3: Secondary structural elements mediate differential tolerance to deletions versus substitutions. (a) The distribution of secondary structural elements genome-wide (blue bars), among $D^{-}S^{+}$ sites (red bars), and among $D^{+}S^{-}$ sites (green bars). **(b)** Difference in average ESM score for each amino acid substitution pair, comparing scores within versus outside helices against the difference in volume between wildtype and mutant residue. Substitutions to larger (left) and smaller (right) residues are shown separately. Orange lines indicate linear fits, with slope and p-values reported. Each dot corresponds to a specific substitution (i.e. F>W) within versus outside helices. **(c)** Comparison of substitution effects versus insertion effects within and outside helices across 1,632 proteins. See **Methods** for scoring details **(d)** Change in experimentally-measured stability from insertions and deletions, normalized by average change in stability from substitutions, within secondary structural elements (N=11,777 sites)[10].

Sequence context mediates $D^{+}S^{-}$ differential tolerance

Unlike sites that are deletion-*intolerant* and substitution-*tolerant* ($D^{-}S^{+}$ red quadrant, **Fig. 2b**), sites that tolerate deletions but not substitutions ($D^{+}S^{-}$, green quadrant, **Fig. 2b**) do not segregate as unambiguously by secondary structure (**Fig. 3a**). In order to understand why certain sites are deletion-tolerant and substitution-intolerant, we identified sequences enriched for this property.

Among our examples, we observed many regions composed of chemically similar (or identical) amino acids. For example, the BRD2 protein (UniProt accession P25440) contains a 30-residue region, with 12 sites meeting the D^+S^- criteria. The amino acids in this region are small and polar, with 21 serines and four aspartic acids. We therefore considered whether D^+S^- sites were more likely to emerge within homogenous sequence contexts; as in the above example, one deletion from a string of many small, polar residues might be less damaging than substituting in a large, hydrophobic residue.

To test this, we needed to quantify the homogeneity of a given sequence. We first used BLOSUM62 substitution scores, which can be used to approximate a similarity measure between neighboring residues. Under this definition of homogeneity, we find that D^+S^- sites are more similar to their neighboring residues than both D^-S^+ sites and random positions sampled across all proteins (**Fig. 4a**, **Supplementary Figs. S8** and **S9**). As in our secondary structure analysis, we find that this property is clarified by the joint analysis of deletions alongside substitutions. (**Supplementary Fig. S10)**.

As an alternative approach, we used amino acid clusters based on shared properties[32–34] to measure the homogeneity of a given sequence. Consistent with our prior result, we find that the sequence context around D^+S^- sites is more homogenous relative to the contexts of both D^-S^+ sites and a random background (**Fig. 4b**). This finding is highly robust to the choice of amino acid clustering, the amino acid distribution of the background set, and the length of the sequence context (**Supplementary Figs. S11-13**).

Our analysis suggests that D^+S^- sites are embedded in regions that are robust to small perturbations in length (i.e., deletions) but not perturbations in context. With this understanding,

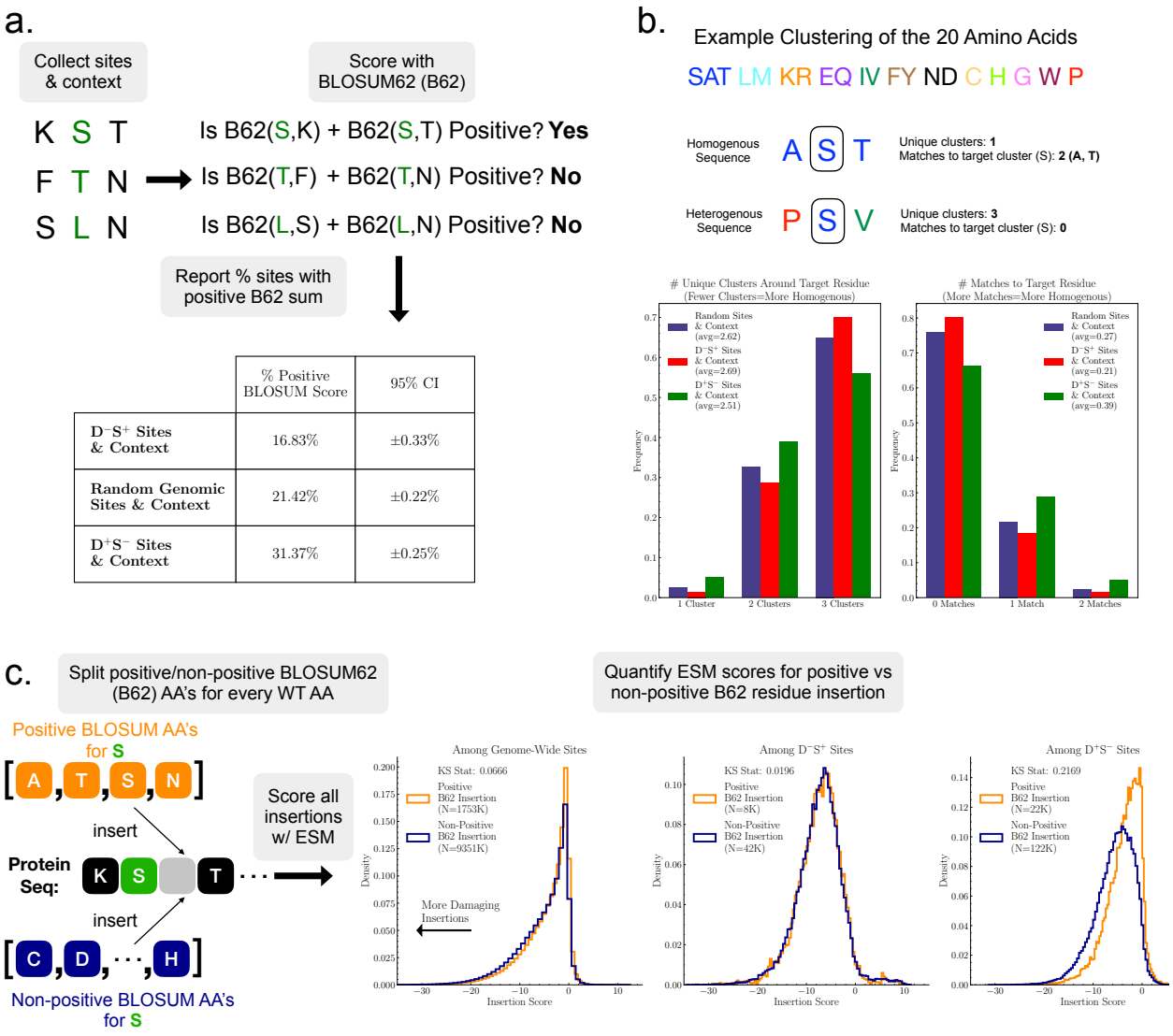
we also anticipated that these regions should be insertion-tolerant, especially when the inserted amino acid is concordant with the local sequence context. For amino acids that are discordant with the sequence context, we expected a more deleterious distribution of effects.

To explore this claim, we returned to our insertion scanning dataset. At every site, the effect of inserting any amino acid was evaluated with ESM. To identify insertions that are concordant with the sequence context, we again used BLOSUM62 scores: a positive BLOSUM62 score between the inserted residue and its N-terminal neighbor denotes a concordant insertion, while a non-positive BLOSUM62 score denotes a discordant insertion. In alignment with our expectations, we observe that D^+S^- sites display differential tolerance to insertions. This differential tolerance is mediated by the concordance (or lack thereof) between the inserted residue and the sequence context (**Fig. 4c**).

Additionally, we considered whether differentially tolerant sites are more likely to participate in ligand binding. We find D^+S^- are more likely to be ligand binding sites than D^-S^+ sites or the genome-wide background rate (**Supplementary Fig. S14**). This is perhaps consistent with our results on homogenous sequence contexts, given that ligands can have strong sequence preferences for binding. ATP, for example, has a binding preference for positively charged residues Arg, Lys, and His.[35]

In summary, we find homogenous sequence context, as measured by BLOSUM62 substitution scores and amino acid clusters, to be a distinctive feature of sites that can tolerate deletions but not substitutions (D^+S^-). These regions are robust to small perturbations in length but not perturbations in amino acid properties. Our insertion scan data further demonstrates this point, as D^+S^- sites are in regions that harbor differential tolerance to insertions, mediated by the identity

of the inserted residue. This property is not observed across $D^{-}S^{+}$ sites nor among the full insertion dataset.



	% Positive BLOSUM Score	95% CI
$D^{-}S^{+}$ Sites & Context	16.83%	±0.33%
Random Genomic Sites & Context	21.42%	±0.22%
$D^{+}S^{-}$ Sites & Context	31.37%	±0.25%

Figure 4: Homogenous sequence contexts mediate differential tolerance to deletions versus substitutions. (a) BLOSUM62 (B62) scores for sequence homogeneity. The similarity of a residue and its neighbors is scored with B62. The sign of the sum of B62 scores is used as a proxy of sequence homogeneity. Sequence homogeneity score shown for $D^{-}S^{+}$ sites, random sites from the proteome (N=135,025), and $D^{+}S^{-}$ sites. **(b)** MMSEQS12 amino acid clustering and scoring of a site's context. Results for sites (and context) randomly selected from the genome and among sites with differential tolerance. *(figure caption continued on next page)*

(figure caption continued from previous page) **(c)** Insertion scan dataset is split between insertions that are similar to its wildtype neighbor (positive B62 score, orange distribution) and all other residues (non-positive B62 score, navy distribution). Insertion scores are reported within these two cases for the full insertion dataset and among sites in the insertion dataset that display differential tolerance. Kolmogorov-Smirnov test statistic is reported.

Differential tolerance can impact the interpretation of DMS experiments

Of the 20,298 human proteins considered in our *in-silico* analysis, 17,631 (nearly 87%) contain at least one site that meets our criteria for differential tolerance. Thus, for the vast majority of human proteins, considering substitutions independently of deletions may generate misleading conclusions on certain sites' overall mutational tolerance.

As a consequence of this finding, and a strong bias towards substitutions in existing experimental datasets, we hypothesized that results from deep mutational scanning could benefit from the joint consideration of deletions alongside substitutions. To substantiate this, we first re-analyzed changes in folding stability from substitutions and deletions from Tsuboyama et al. (2023), where we identified differential tolerance among the sites assayed (**Fig. 2b**). Tsuboyama et al. used a high-throughput cDNA display proteolysis assay to measure the effect of substitutions and deletions across thousands of sites within natural and *de novo* designed protein domains.

We used ESM-1b to score all sites from Tsuboyama et al., totaling 11,777 substitution-deletion pairs. When compared against our differentially tolerant thresholds, we find a small number of sites satisfying either the $D^{-}S^{+}$ or $D^{+}S^{-}$ criteria (**Fig. 5a)**. We then re-incorporated the experimental data to validate the differentially tolerant predictions. In general, we find the experimental results are consistent with our computational predictions: $D^{-}S^{+}$ sites are significantly more deletion intolerant than $D^{+}S^{-}$ sites, and $D^{+}S^{-}$ sites are significantly more substitution intolerant than $D^{-}S^{+}$ sites (**Fig 5b**). Because the Tsuboyama et al. data consists only

of short human and non-human functional domains, it is not directly comparable to our genome-wide data, which is composed of full-length human proteins. Moreover, the relationship between protein fitness (proxied with ESM) and protein stability (measured by Tsuboyama et al.) is not one-to-one.[36,37] Nonetheless, the alignment between our computational predictions and experimental results suggests that differential tolerance exists in DMS data, and that ESM can effectively indicate its presence.

We observed many instances of agreement between our computationally-predicted differentially tolerant sites and the corresponding experimental data. For example, the J-domain of Tid1[38] (PDB: 6IWS) has an alanine at position 48 that is predicted (via ESM) to be $D^{-}S^{+}$. The experimental data is in agreement with this prediction: 6IWS A48 is more deletion-*intolerant* than 90% of deletions and more substitution-*tolerant* than 83% of substitutions in Tsuboyama et al. Consistent with our results on $D^{-}S^{+}$ sites, A48 is contained within an alpha helix. Structural prediction with ESMFold[16] suggests that A48 plays an important role in maintaining this helix, as its deletion creates a new unstructured region and results in an overall decrease in model confidence (**Fig. 5c**). However, no substitution at this site has the same structural consequence (**Fig. 5d**). While validating this structural perturbation is beyond the scope of the paper, the ESMFold prediction provides a mechanism for the site's observed differential tolerance.

We also identified cases of agreement between experimental measurements and computational predictions among $D^{+}S^{-}$ sites. For example, I36 of the peptidoglycan binding domain[39] (PDB: 2MKX) is predicted by ESM to be $D^{+}S^{-}$. Experimental results suggest a similar tendency, with 2MKX I36 more deletion-*tolerant* than 83% of all deletions and more substitution-*intolerant* than 97% of all substitutions in Tsuboyama et al. I36 is embedded within a homogenous

sequence context, neighbored by hydrophobic residues: an N-terminal Leu, and a C-terminal Phe and Val (**Fig. 5d**).[40] We reasoned that this sequence context drives the differential tolerance observed at I36: this part of the domain must be hydrophobic, but it is robust to a small decrease in length. Consistent with this claim, we observe a clear relationship between residue hydrophobicity and substitution effect at 2MKX I36. In particular, while the site is highly intolerant of substitutions overall, this intolerance is most pronounced for substitutions to a hydrophilic residue. Substitutions that maintain the hydrophobicity of the site are more neutral in effect (**Fig. 5e**).

Overall, these results underscore the value of considering substitutions alongside deletions in studies of protein fitness. Evidence of differential tolerance can be found in existing experimental datasets. If either deletion or substitution effects were excluded from these analyses, an incomplete understanding of mutational tolerance would be drawn for sites like 6IWS A48 and 2MKX I36. Moreover, we have shown that our genome-wide thresholds and ESM scores can be used to predict differential tolerance in protein sequences of interest.

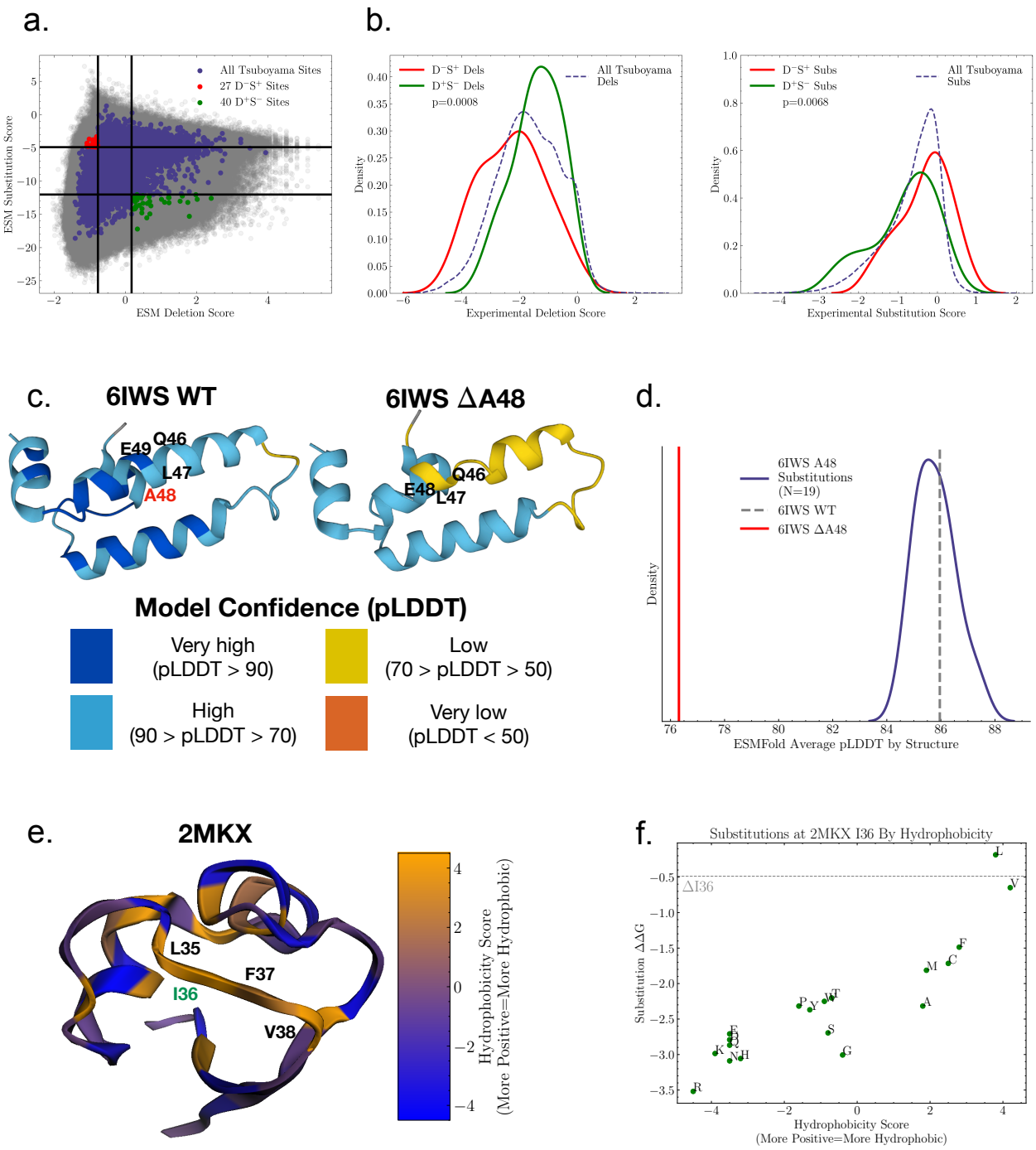


Figure 5: Differential tolerance can be predicted in mutational scanning data. *(figure caption continued on next page)*

(figure caption continued from previous page) **(a)** All sites from Tsuboyama et al. scored via ESM on their deletion and substitution tolerance. Differentially tolerant sites were identified via comparison to our genome-wide thresholds. **(b)** Distribution of experimental deletion scores (left) or substitution scores (right) among differentially tolerant sites identified in Tsuboyama et al. Full distribution of experimental deletions (left) or substitutions (right) from Tsuboyama et al. given by the dotted line. P-values correspond to the t-test for difference in means between the two differential tolerance conditions. **(c)** ESMFold structures 6IWS WT (left) and 6IWS ΔA48 (right) colored by model confidence (pLDDT). **(d)** ESMFold pLDDT (average per structure) for 19 substitutions at 6IWS A48 (blue distribution), wildtype 6IWS (dashed grey line), and 6IWS ΔA48 (solid red line). **(e)** PDB structure of 2MKX colored by hydrophobicity. I36 and sequence context highlighted. **(f)** Substitution ΔΔG at 2MKX I36 by hydrophobicity of the mutant residue. ΔI36 ΔΔG is shown by the dotted line.

Discussion

Deletions are a critical class of mutations in protein evolution, engineering, and human genetics. Despite this, their genome-wide effects remain poorly understood, particularly relative to substitutions. To further our understanding of the relationship between deletions and substitutions, we used the ESM-1b protein language model to predict the effects of all amino acid deletions in the human proteome.

We then used ESM-1b to compare deletion versus substitution tolerance genome-wide. The scale of this analysis allowed us to identify hundreds of thousands of sites that display differential tolerance. The existence of these sites suggests that substitution effects, while correlated with deletion effects in aggregate, are insufficient to fully capture the tendencies of deletion intolerance.

Among sites that are deletion-intolerant and substitution-tolerant ($D^{-}S^{+}$), we find secondary structural elements to be an important predictive feature. In particular, sites contained within alpha-helices are strongly overrepresented among $D^{-}S^{+}$ sites relative to their background frequency in the human proteome. On the other hand, sites that are deletion-tolerant but

substitution-intolerant (D^+S^-) tend to be found within more homogenous sequence contexts than D^-S^+ sites and random sites from the proteome. Our results provide genome-wide evidence for differential tolerance between substitutions and deletions across *in-silico* and experimental datasets and establish protein language models as a powerful tool for biological hypothesis generation and analysis.

Our results pose numerous areas for further exploration. To begin, we identify certain features associated with differential tolerance, namely alpha-helices and homogenous sequence context. However, not every site that displays differential tolerance possesses these features. Relatedly, not every site with these features displays differential tolerance. Resolving these ambiguities will create a more nuanced view of why differential tolerance does (and does not) occur.

As demonstrated, the principle of differential tolerance can help to further refine clinical annotations. In particular, existing clinical guidelines draw an explicit link between amino acid conservation and the inferred deleteriousness of length-modifying mutations.[41] While refining such guidelines is beyond the scope of this study, we argue that differential tolerance may challenge the direct transfer of clinical annotations between a substitution and a deletion at the same site.

Lastly, our results demonstrate the power of protein language models to proxy an experimental system. This class of models has already demonstrated highly accurate predictive performance across relevant biological domains. Here, we demonstrate their utility in answering second-order biological questions, such as the comparison between different types of mutations. Our analysis illustrates the potential for these tools to supplement limited experimental data to generate new biological hypotheses.

Code and Data Availability

Deletion and average substitution scores from ESM-1b can be downloaded at https://huggingface.co/spaces/ntranoslab/diff-tol. Other datasets used for analysis can be downloaded at our GitHub https://github.com/ntranoslab/diff-tol.

Methods

Accessing ESM ESM-1b was accessed pretrained from https://github.com/facebookresearch/esm using the `esm.pretrained.esm1b_t33_650M_UR50S` module.

Scoring substitutions with ESM We scored substitutions with ESM via the log-likelihood ratio (LLR), which compares the log-likelihood of the mutant residue against the log-likelihood of the reference residue. Code for scoring substitutions via the LLR can be found at https://github.com/ntranoslab/esm-variants/blob/main/esm_score_missense_mutations.py. Because each site has 19 substitution scores but only one deletion score, we aggregated substitution effects at the site level to compare the measurements one-to-one. Among the different aggregation functions tested, we found that a site's average LLR most accurately predicts ClinVar annotations compared to the site's median, max, and min LLR. As such, we chose the average LLR to represent a site's substitution tolerance. Mathematically, the LLR is computed as $\log\left(\frac{P\left(x_{i,mut}|s\right)}{P\left(x_{i,wt}|s\right)}\right)$ where s is the wildtype sequence, $x_{i,wt}$ is the wildtype residue at position i in sequence s, and $x_{i,mut}$ is the mutant residue at position i. The site's

average LLR is $\frac{1}{19}\sum_{mut \not\equiv wt} \log\left(\frac{P\left(x_{i,mut}|s\right)}{P\left(x_{i,mut}|s\right)}\right)$, or the average LLR across all 19 possible substitutions.

PLLR framework for scoring mutations with ESM The pseudo log-likelihood ratio (PLLR) is defined as the ratio of the pseudo-log-likelihoods of the *entire* mutant sequence and the *entire* wildtype sequence, as estimated by ESM[3]. All deletions and insertions in this study were scored with the PLLR. We also used the PLLR metric for substitutions in **Fig. 3c** and **Supplementary Fig. S4.** Code for scoring mutations via the PLLR can be found at https://github.com/ntranoslab/esm-variants/blob/main/esm_score_multi_residue_mutations.py. Mathematically, the PLLR is computed as $PLL(s^{mut}) - PLL(s^{wt})$, where $PLL(s) = \sum_i \log\left(P\left(x_i = s_i | s_{/i}\right)\right)$, where s is the given sequence, and $P\left(x_i = s_i | s_{/i}\right)$ is the probability the model puts on the wildtype residue at position i in the context of sequence s masked at position i. For computational efficiency, we use the "single pass" PLL, given by $PLL(s) = \sum_i \log\left(P\left(x_i = s_i | s\right)\right)$ which does not require masking at every position.[3,42]

Comparing ESM substitution and deletion scores Because substitution tolerance is represented by the average LLR, and deletion tolerance is represented by the PLLR, we cannot directly compare substitution scores against deletion scores. To address this, we partition the deletion data and substitution data separately, which avoids any direct comparison across the two distributions.

Scoring long sequences with ESM To overcome ESM-1b's hard-coded 1,022 sequence length limit, we re-employed our existing tiling framework.[3]

Visualizing the distribution of deletion effects Prior research indicates that the absolute PLLR (aPLLR) is the most accurate predictor of benign/pathogenic clinical annotations[3]. As such, we conducted all deletion analyses in Fig. 3, including the data partitioning, using the aPLLR. To aid with visualization, however, we *plot* the aPLLR's on the -log10 scale in **Fig. 2a** and **Fig. 2c**. This facilitates the visual inspection of deletion effects across the aPLLR range.

Accessing ClinVar annotations for substitutions and deletions The median ESM LLR of ClinVar benign and ClinVar pathogenic substitutions were taken from https://github.com/ntranoslab/esm-variants/blob/main/benchmarks/ClinVar_gnomAD_benchmark_with_predictions.csv. For deletions, we used https://github.com/ntranoslab/esm-variants/blob/main/benchmarks/ClinVar_indel_benchmark_with_predictions.csv and filtered to only deletions (removing insertions and del-ins).

Secondary structural element prediction All AlphaFold-predicted human protein structures were extracted from https://www.uniprot.org/proteomes/UP000005640. Secondary structural elements were assigned using the `Bio.PDB.DSSP` module from the BioPython library (https://biopython.org/docs/1.75/api/Bio.PDB.DSSP.html). We grouped DSSP annotations into three categories: helices (H, G, I), strands (E, B), and loops (S, T). All figures analyzing properties of helices (**Figs. 3a-c**) used these predictions from AlphaFold and DSSP. However, for **Fig. 3b** only, we restricted the definition of helices to be those with an average AlphaFold

pLDDT > 0.9 to remove noise from lower-confidence helices. All other analyses used all helices regardless of pLDDT.

Insertion and substitution scanning dataset We selected 1,632 proteins whose structures were approximately 50% alpha-helix, as estimated by DSSP. All proteins selected correspond to canonical transcripts from the AlphaFold database. For each site, we used the PLLR metric to estimate the effect of both (a) inserting a mutant residue to the right (C-terminal side) of the wildtype residue, and (b) substituting the wildtype residue to the mutant residue. In **Fig. 3c**, we excluded synonymous substitutions and insertions of the same amino acid as wildtype to avoid biasing the results towards insertions being more damaging. For this dataset, we consider the non-averaged insertion and substitution effects.

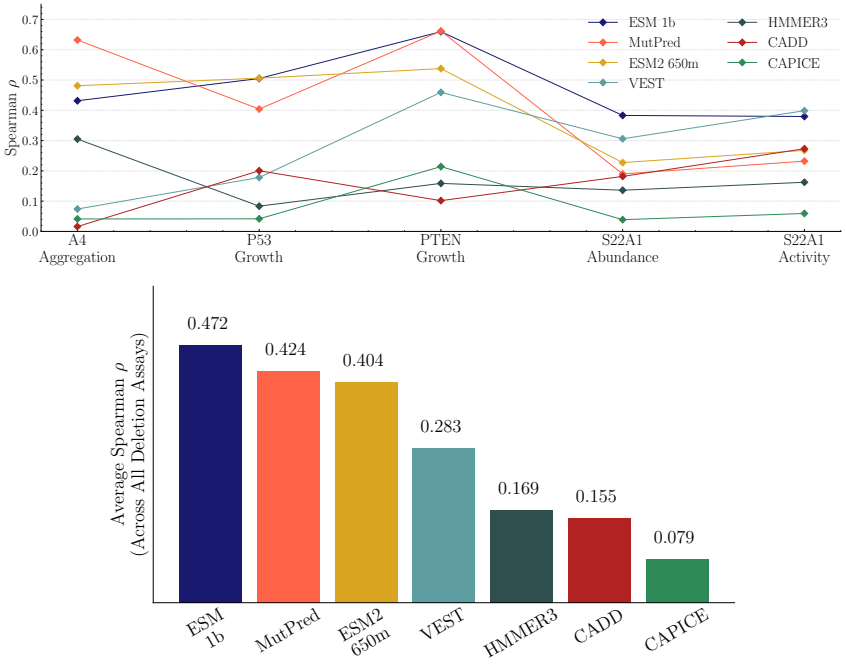
Analysis of experimental stability measurements DMS data was downloaded from Tsuboyama et al. (https://zenodo.org/records/7992926). The average substitution ΔΔG was computed per position. Variants not mapping to a PDB identifier or absent of any substitution, insertion, or deletion ΔΔG values were removed. For duplicate PDB identifiers with a single amino acid substitution (i.e., 1BK2.pdb_F47A and 1BK2.pdb_L5S), we select whichever sequence has the most observations and drop all others, to avoid redundancy in the data.

- DMS ΔΔG data: processed_K50_dG_datasets/Tsuboyama2023_Dataset2_Dataset3_20230416.csv
- Secondary structure data: Data_tables_for_figs/dG_site_feature_Fig3.csv.

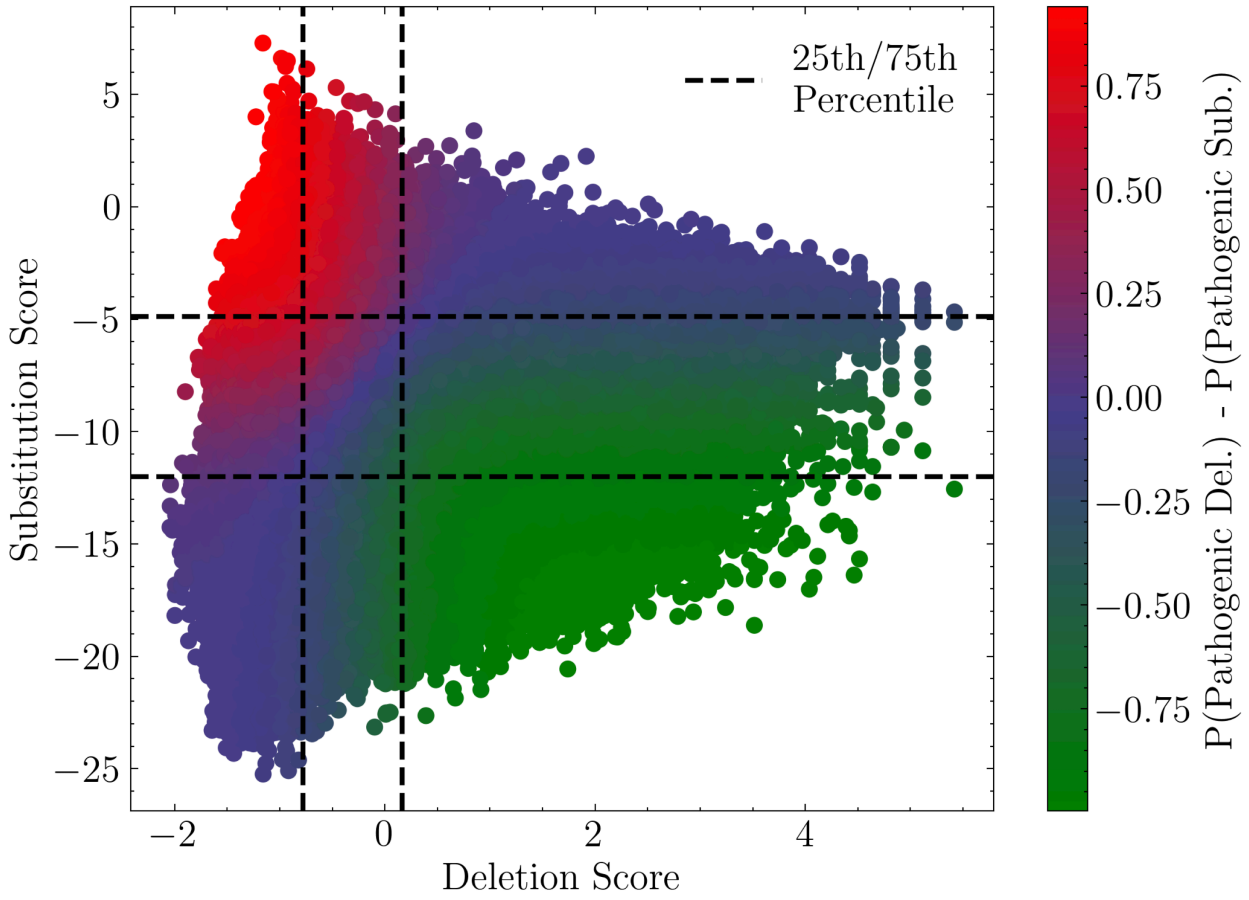
Supplementary Figures

a.

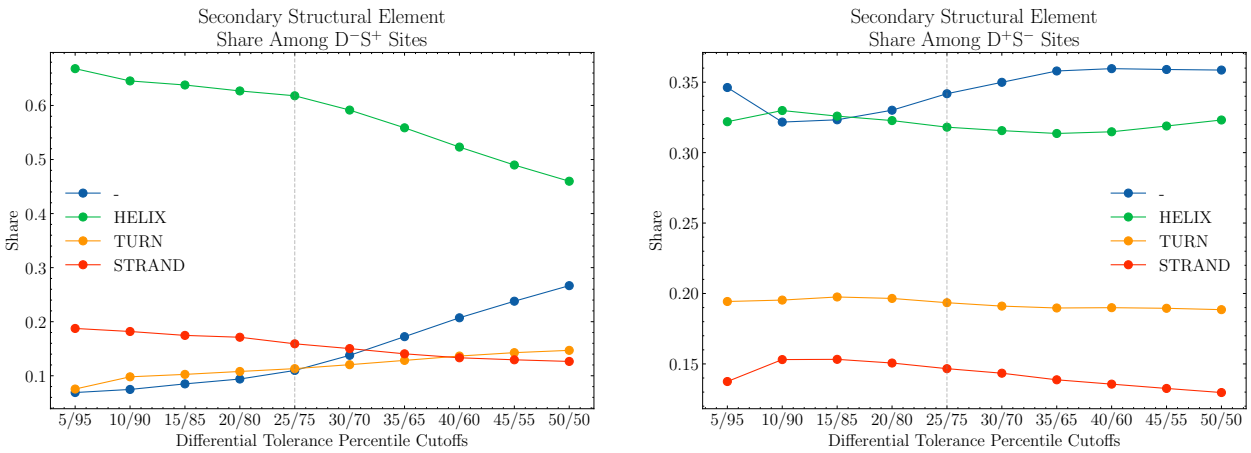
b.



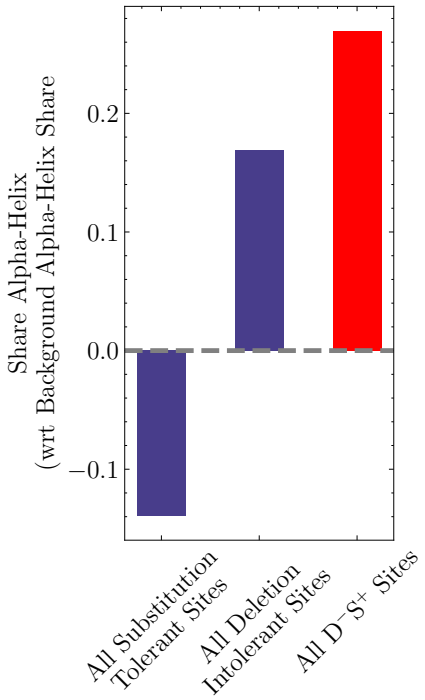
Supplementary Figure S1: Comparison of variant effect prediction methods on predicting deletion assays. *Related to Figure 1.* **(a)** The spearman correlation between the variant effect predictor and the deletion assay. Assays were curated from Protein Gym[6]. **(b)** Average spearman correlation across all assays for each method.



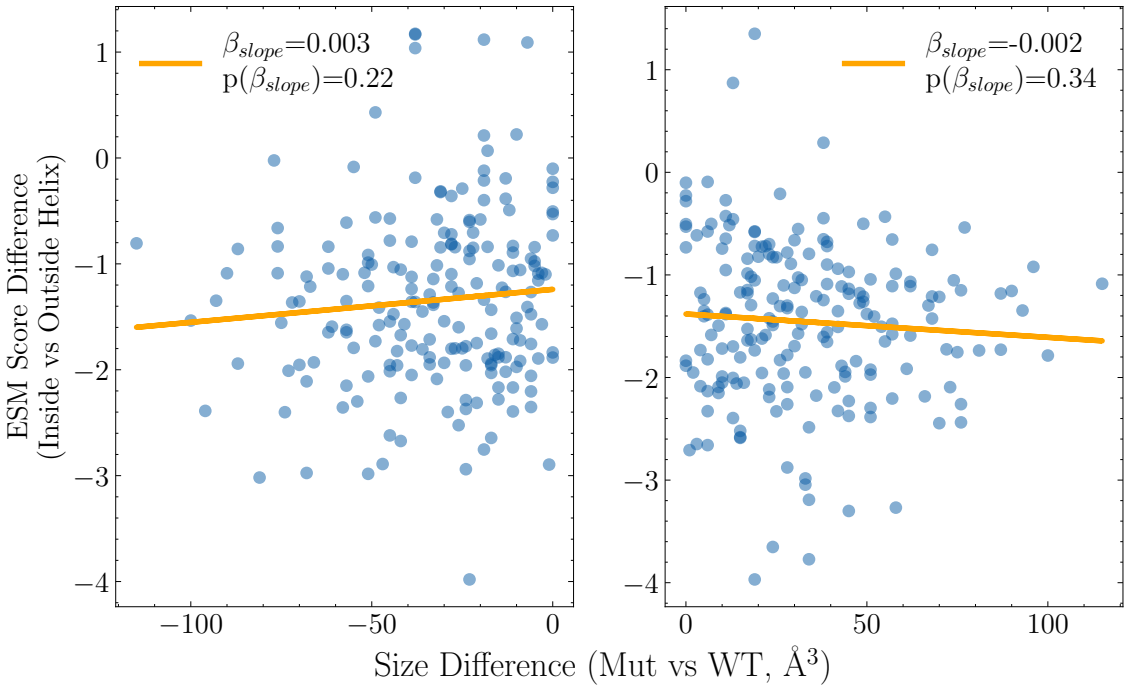
Supplementary Figure S2: Continuous measure of differential tolerance. *Related to Figure 2.* Each site's deletion and substitution scores each were re-scaled between 0 and 1 by training deletion and substitution logistic regression models on ClinVar annotations (benign/pathogenic). The score can be interpreted as the probability that the deletion (or substitution) is pathogenic, given the ESM score. The difference between the rescaled substitution score and deletion score is used to approximate differential tolerance: a difference of 1 implies deletion intolerance and substitution tolerance, and a score of -1 implies deletion tolerance and substitution intolerance. The 25th/75th percentile thresholds are shown for comparison.



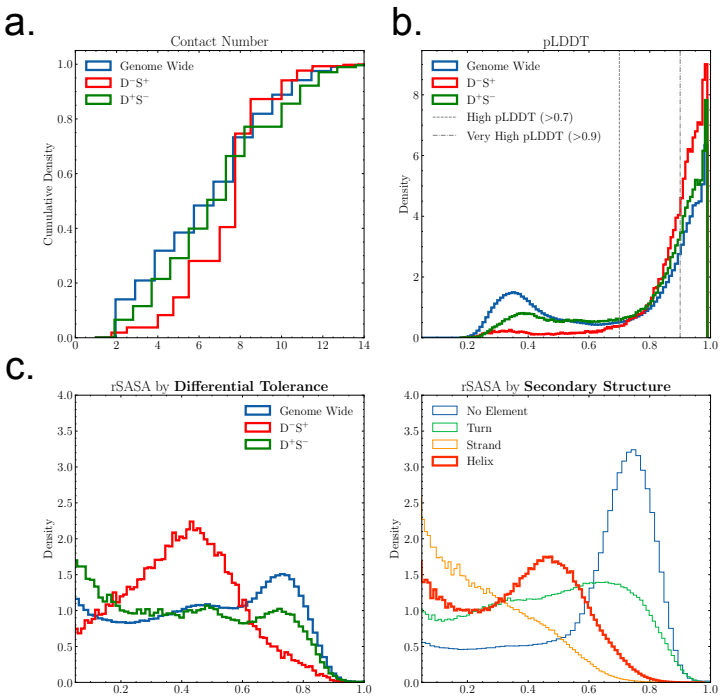
Supplementary Figure S3: Secondary structural element shares in $D^{-}S^{+}$ $D^{+}S^{-}$ quadrants as a function of percentile cutoff. *Related to Figure 3.* Share of sites belonging to each DSSP category in the deletion-intolerant, substitution tolerant ($D^{-}S^{+}$) quadrant or deletion tolerant, substitution-intolerant ($D^{+}S^{-}$) quadrant when varying the percentile cutoff. The 25th/75th percentile cutoff was selected for the main analysis.



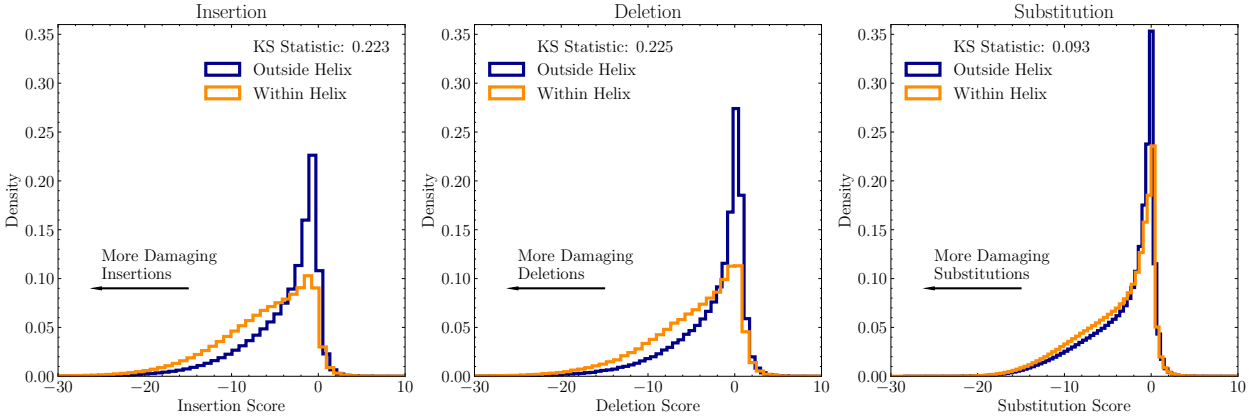
Supplementary Figure S4: Share alpha-helix within different data partitions. *Related to Figure 3*. Relative to the background frequency in the data, the share of alpha-helices among all substitution-tolerant sites, all deletion-intolerant sites, and all $D^{-}S^{+}$ sites, with respect to (wrt) the proteome-wide share of alpha helices.



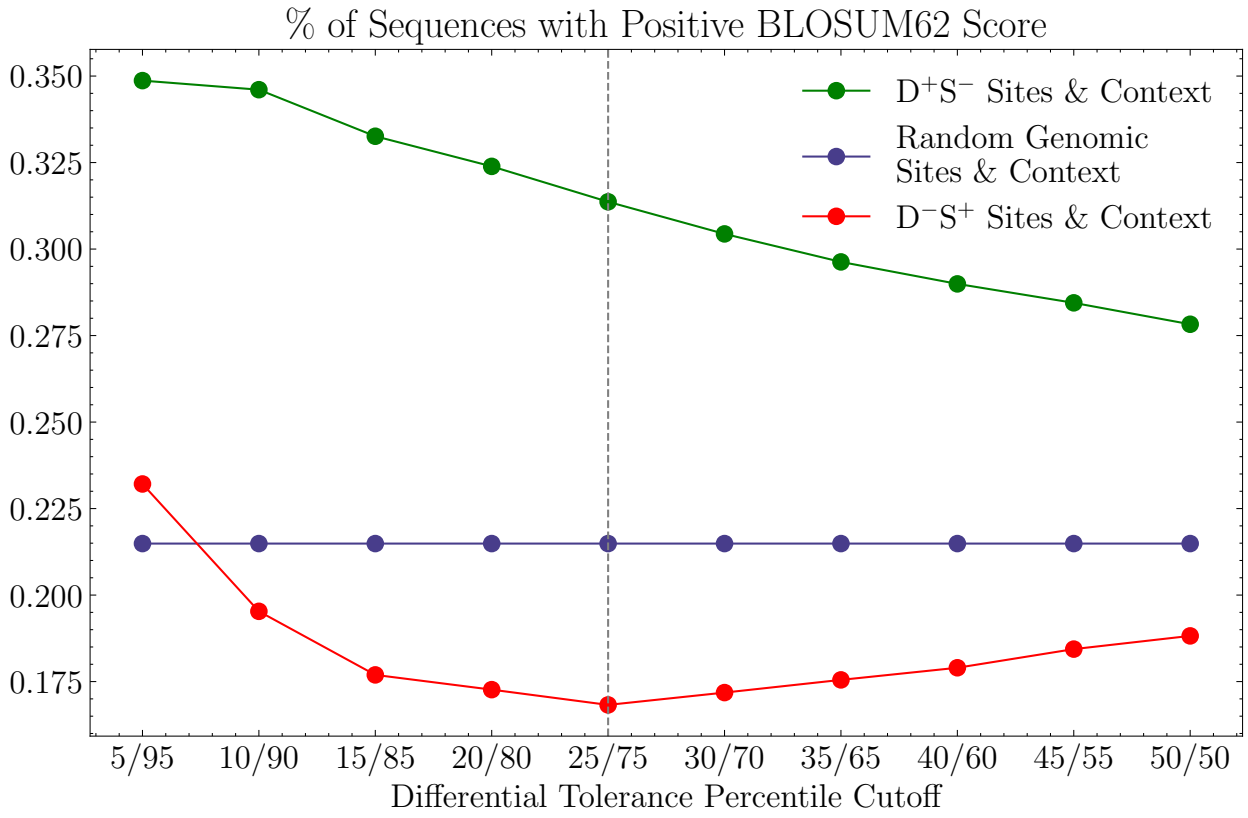
Supplementary Figure S5: Average substitution effect inside versus outside helices by wildtype and mutant size difference. *Related to Figure 3.* For all possible substitution pairs, the average effect of the mutation outside helices versus inside helices is compared against the pairs' difference in size. In this analysis, no pLDDT restriction is put on helix inclusion. The slope coefficients and coefficient p-values for the lines of best fit are reported.



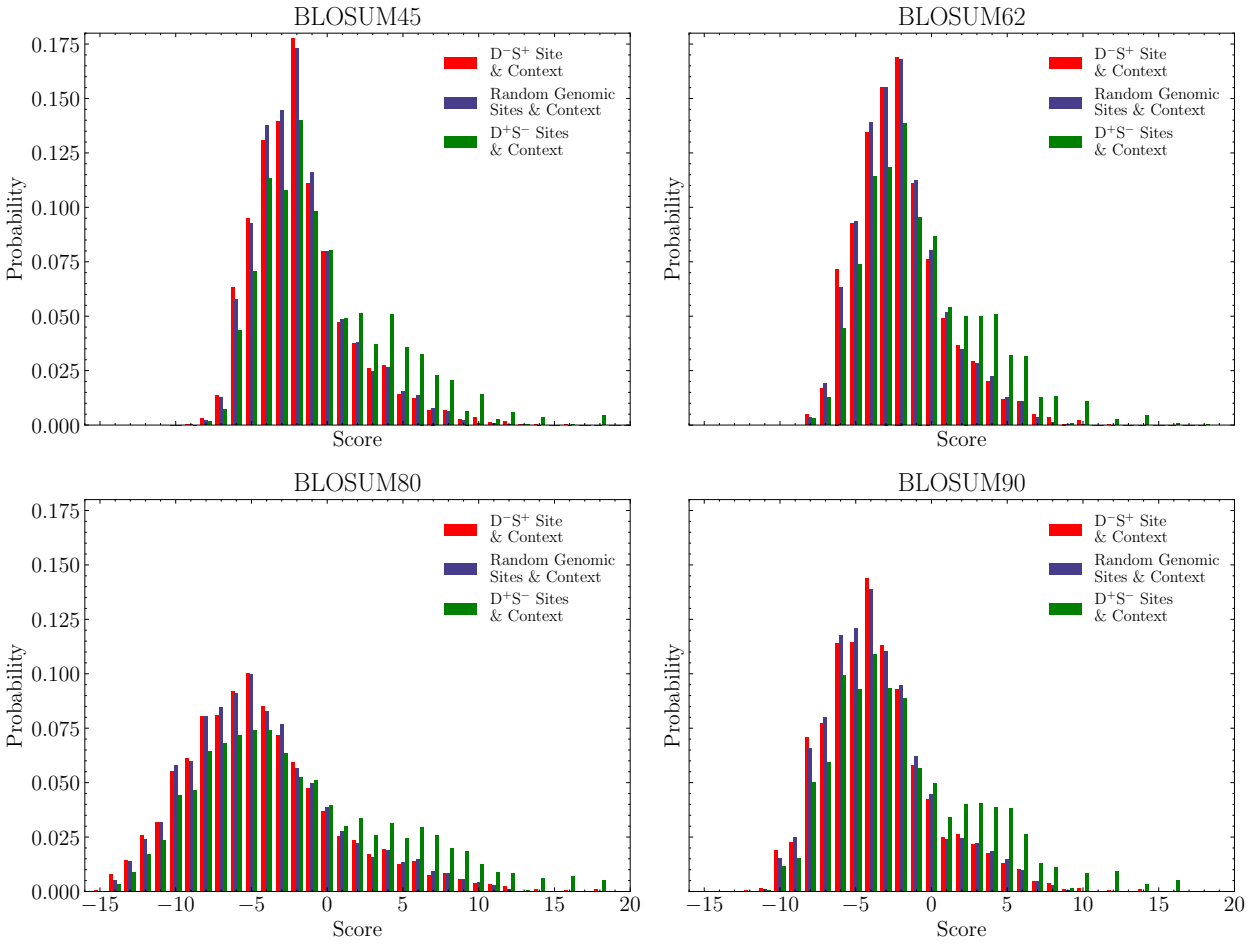
Supplementary Figure S6: Structural correlates of differential tolerance. *Related to Figure 3.* **(a)** Cumulative density of contact number (<4Å) within differential tolerance condition. **(b)** Distribution of AlphaFold pLDDT by differential tolerance condition. **(c)** rSASA by differential tolerance condition (right) and by secondary structure annotation (left).



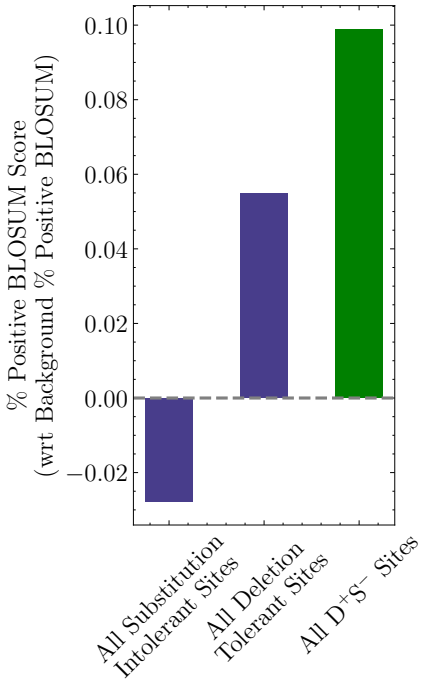
Supplementary Figure S7: Tolerance to insertions, deletions, and substitutions within versus outside helices. *Related to Figure 3.* Insertions, deletions, and substitutions were scored with the ESM PLLR metric for 1,632 proteins. The effect of the mutation within versus outside a helix is shown.



Supplementary Figure S8: Percent of sequences with positive BLOSUM62 (B62) scores as a function of percentile cutoff. *Related to Figure 4.* Approximating sequence homogeneity with B62 scores among random genomic sequences and within each quadrant when varying the percentile cutoff used to construct the quadrants.



Supplementary Figure S9: Distribution of BLOSUM sequence sums for multiple substitution matrices. *Related to Figure 4.* Scoring sequence homogeneity using BLOSUM substitution matrices. A positive value indicates a more homogenous sequence context.

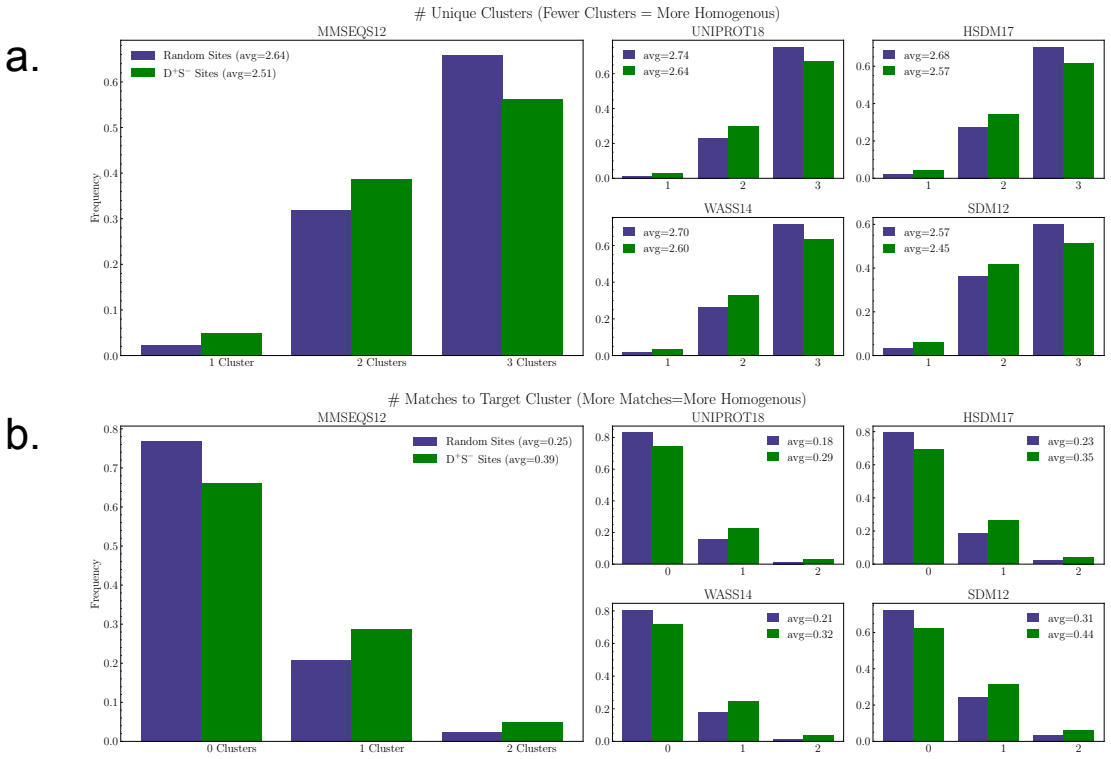


Supplementary Figure S10: Percent positive BLOSUM score within different data partitions. *Related to Figure 4.* Relative to the background share in the data, the percent of sites and context with a positive BLOSUM score among all substitution-intolerant sites, all deletion-tolerant sites, and all $D^{+}S^{-}$ sites.

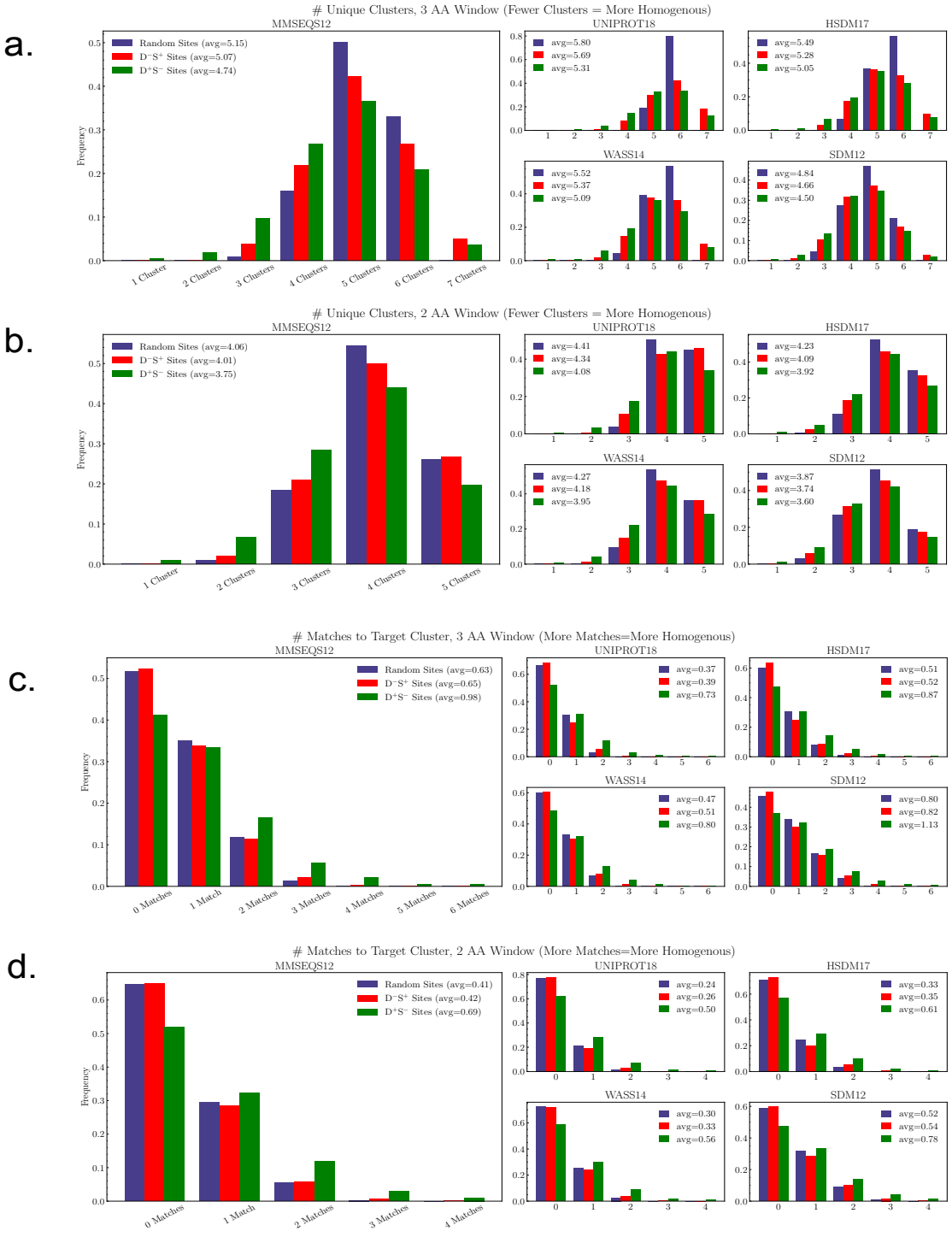
a.

b.

Supplementary Figure S11: Scoring sequence homogeneity across an expanded set of amino acid clusterings. *Related to Figure 4.* **(a)** Counting the number of unique clusters represented in a sequence window among genome-wide sites (N=135,025, blue bars) and all sites in the $D^{-}S^{+}$ and $D^{+}S^{-}$ quadrant (red and green bars, respectively) for an expanded set of amino acid clustering approaches. The number of clusters is given in the cluster name. Average values reported under the MMSEQS12 clustering. **(b)** Counting the number of cluster matches to the target residue in a sequence window among genome-wide sites (blue bars) and all sites in the $D^{-}S^{+}$ and $D^{+}S^{-}$ quadrant (red and green bars, respectively) for an expanded set of amino acid clustering approaches. The number of clusters is given in the cluster name. Average values reported under the MMSEQS12 clustering.

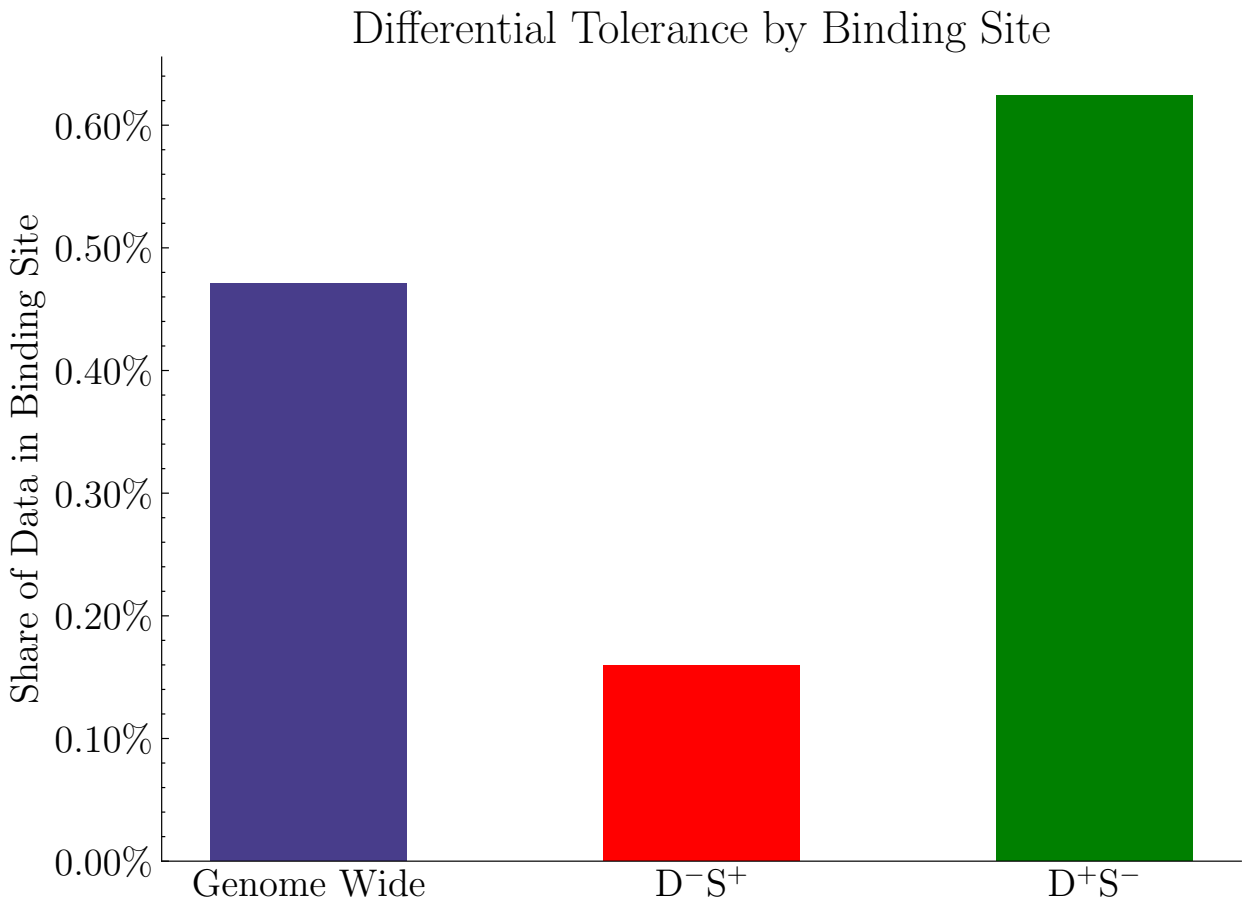


Supplementary Figure S12: Scoring sequence homogeneity with a dataset matched on amino acid frequency. *Related to Figure 4.* Sequence homogeneity scores between sites in the D^+S^- quadrant versus a genome-wide set. The genome-wide set was redrawn to match the frequency of amino acids in the D^+S^- quadrant. This adjusts for the tendency of certain residues (i.e., glycine, proline, tryptophan) to be in their own cluster across alphabets, which makes sequences that contain them appear more heterogeneous. **(a)** Number of unique clusters, and **(b)** number of cluster matches. Average values reported under the MMSEQS12 clustering.



Supplementary Figure S13: Scoring sequence homogeneity across a wider sequence range. *Related to Figure 4.* **(a)** Number of unique clusters represented around a three amino acid window (in either direction) of the target residue. *(figure caption continued on next page)*

(figure caption continued from previous page) **(b)** Number of unique clusters represented around a two amino acid window (in either direction) of the target residue. **(c)** Matches to target residue around a three amino acid window (in either direction) of the target residue. **(d)** Matches to target residue around a two amino acid window (in either direction) of the target residue. Average values reported under the MMSEQS12 clustering.



Supplementary Figure S14: Differential tolerance by binding site. *Related to Figure 4.* Ligand binding data from ChEBI used to identify the percentage of sites in each condition that participate in ligand binding.

References

1. Rehm, H.L., and Fowler, D.M. (2019). Keeping up with the genomes: Scaling genomic variant interpretation. Genome Medicine *12*. https://doi.org/10.1186/s13073-019-0700-4.

2. Horne, J., and Shukla, D. (2022). Recent Advances in Machine Learning Variant Effect Prediction Tools for Protein Engineering. Ind. Eng. Chem. Res. *61*, 6235–6245. https://doi.org/10.1021/acs.iecr.1c04943.

3. Brandes, N., Goldman, G., Wang, C.H., Ye, C.J., and Ntranos, V. (2023). Genome-wide prediction of disease variant effects with a deep protein language model. Nat Genet *55*, 1512–1522. https://doi.org/10.1038/s41588-023-01465-0.

4. Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J.K., Brock, K., Gal, Y., and Marks, D.S. (2021). Disease variant prediction with deep generative models of evolutionary data. Nature *599*. https://doi.org/10.1038/s41586-021-04043-8.

5. Esposito, D., Weile, J., Shendure, J., Starita, L.M., Papenfuss, A.T., Roth, F.P., Fowler, D.M., and Rubin, A.F. (2019). MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. Genome Biol *20*, 223. https://doi.org/10.1186/s13059-019-1845-6.

6. Notin, P., Kollasch, A.W., Ritter, D., Van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Weitzman, R., Frazer, J., et al. (2023). ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction (Synthetic Biology) https://doi.org/10.1101/2023.12.07.570727.

7. Macdonald, C.B., Nedrud, D., Grimes, P.R., Trinidad, D., Fraser, J.S., and Coyote-Maestas, W. (2023). DIMPLE: deep insertion, deletion, and missense mutation libraries for exploring protein variation in evolution, disease, and biology. Genome Biol *24*, 36. https://doi.org/10.1186/s13059-023-02880-6.

8. Savino, S., Desmet, T., and Franceus, J. (2022). Insertions and deletions in protein evolution and engineering. Biotechnology Advances *60*, 108010. https://doi.org/10.1016/j.biotechadv.2022.108010.

9. Zhang, Z., Wang, J., Gong, Y., and Li, Y. (2018). Contributions of substitutions and indels to the structural variations in ancient protein superfamilies. BMC Genomics *19*, 771. https://doi.org/10.1186/s12864-018-5178-8.

10. Tsuboyama, K., Dauparas, J., Chen, J., Laine, E., Mohseni Behbahani, Y., Weinstein, J.J., Mangan, N.M., Ovchinnikov, S., and Rocklin, G.J. (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.

11. Topolska, M., Beltran, T., and Lehner, B. (2023). Deep indel mutagenesis reveals the impact of amino acid insertions and deletions on protein stability and function (Genomics) https://doi.org/10.1101/2023.10.06.561180.

12. Seuma, M., Lehner, B., and Bolognesi, B. (2022). An atlas of amyloid aggregation: the impact of substitutions, insertions, deletions and truncations on amyloid beta fibril nucleation. Nat Commun *13*, 7084. https://doi.org/10.1038/s41467-022-34742-3.

13. Miton, C.M., and Tokuriki, N. (2023). Insertions and Deletions (Indels): A Missing Piece of the Protein Engineering Jigsaw. Biochemistry *62*, 148–157. https://doi.org/10.1021/acs.biochem.2c00188.

14. Gonzalez, C.E., Roberts, P., and Ostermeier, M. (2019). Fitness Effects of Single Amino Acid Insertions and Deletions in TEM-1 β-Lactamase. Journal of Molecular Biology *431*, 2320–2330. https://doi.org/10.1016/j.jmb.2019.04.030.

15. Larsen-Ledet, S., Lindemose, S., Panfilova, A., Gersing, S., Suhr, C.H., Genzor, A.V., Lanters, H., Nielsen, S.V., Lindorff-Larsen, K., Winther, J.R., et al. (2025). Systematic characterization of indel variants using a yeast-based protein folding sensor. Structure *33*, 262-273.e6. https://doi.org/10.1016/j.str.2024.11.017.

16. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science *379*, 1123–1130. https://doi.org/10.1126/science.ade2574.

17. Lafita, A., Gonzalez, F., Hossam, M., Smyth, P., Deasy, J., Allyn-Feuer, A., Seaton, D., and Young, S. (2024). Fine-tuning Protein Language Models with Deep Mutational Scanning improves Variant Effect Prediction. Preprint at arXiv.

18. Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. bioRxiv. https://doi.org/10.1101/2021.07.09.450648.

19. Bepler, T., and Berger, B. (2021). Learning the protein language: Evolution, structure, and function. Cell Systems *12*. https://doi.org/10.1016/j.cels.2021.05.017.

20. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences of the United States of America *118*. https://doi.org/10.1073/pnas.2016239118.

21. The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Bye-A-Jee, H., et al. (2023). UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research *51*, D523–D531. https://doi.org/10.1093/nar/gkac1052.

22. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res *44*, D862–D868. https://doi.org/10.1093/nar/gkv1222.

23. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*. https://doi.org/10.1038/s41586-021-03819-2.

24. Kabsch, W., and 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.

25. Jacquier, H., Birgy, A., Le Nagard, H., Mechulam, Y., Schmitt, E., Glodt, J., Bercot, B., Petit, E., Poulain, J., Barnaud, G., 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.

26. Kim, R., and Guo, J. (2010). Systematic analysis of short internal indels and their impact on protein folding. BMC Struct Biol *10*, 24. https://doi.org/10.1186/1472-6807-10-24.

27. Surkont, J., Diekmann, Y., Ryder, P.V., and Pereira-Leal, J.B. (2015). Coiled-coil length: Size does matter. Proteins *83*, 2162–2169. https://doi.org/10.1002/prot.24932.

28. Jackson, E.L., Spielman, S.J., and Wilke, C.O. (2017). Computational prediction of the tolerance to amino-acid deletion in green-fluorescent protein. PLoS ONE *12*, e0164905. https://doi.org/10.1371/journal.pone.0164905.

29. Abrusán, G., and Marsh, J.A. (2016). Alpha Helices Are More Robust to Mutations than Beta Strands. PLoS Comput Biol *12*, e1005242. https://doi.org/10.1371/journal.pcbi.1005242.

30. Woods, H., Schiano, D.L., Aguirre, J.I., Ledwitch, K.V., McDonald, E.F., Voehler, M., Meiler, J., and Schoeder, C.T. (2023). Computational modeling and prediction of deletion mutants. Structure *31*, 713-723.e3. https://doi.org/10.1016/j.str.2023.04.005.

31. Elhanati, Y., Sethna, Z., Marcou, Q., Callan, C.G., Mora, T., and Walczak, A.M. (2015). Inferring processes underlying B-cell repertoire diversity. Phil. Trans. R. Soc. B *370*, 20140243. https://doi.org/10.1098/rstb.2014.0243.

32. Ieremie, I., Ewing, R.M., and Niranjan, M. (2024). Protein language models meet reduced amino acid alphabets. Bioinformatics *40*, btae061. https://doi.org/10.1093/bioinformatics/btae061.

33. Steinegger, M., and Söding, J. (2018). Clustering huge protein sequence sets in linear time. Nat Commun *9*, 2542. https://doi.org/10.1038/s41467-018-04964-5.

34. Liang, Y., Yang, S., Zheng, L., Wang, H., Zhou, J., Huang, S., Yang, L., and Zuo, Y. (2022). Research progress of reduced amino acid alphabets in protein analysis and prediction. Computational and Structural Biotechnology Journal *20*, 3503–3510. https://doi.org/10.1016/j.csbj.2022.07.001.

35. Chauhan, J.S., Mishra, N.K., and Raghava, G.P. (2009). Identification of ATP binding residues of a protein from its primary sequence. BMC Bioinformatics *10*, 434. https://doi.org/10.1186/1471-2105-10-434.

36. Beltran, A., Jiang, X., Shen, Y., and Lehner, B. (2025). Site-saturation mutagenesis of 500 human protein domains. Nature *637*, 885–894. https://doi.org/10.1038/s41586-024-08370-4.

37. Gerasimavicius, L., Liu, X., and Marsh, J.A. (2020). Identification of pathogenic missense mutations using protein stability predictors. Sci Rep *10*, 15387. https://doi.org/10.1038/s41598-020-72404-w.

38. Sim, Dae-Won, Jo, Ku-Sung, Ryu, Gyeong-Seok, Kim, Eun-Hui, and 원형식 (2012). Structural Characterization of the J-domain of Tid1, a Mitochondrial Hsp40/DnaJ Protein. Journal of the Korean Magnetic Resonance Society *16*, 22–33. https://doi.org/10.6564/JKMRS.2012.16.1.022.

39. Mesnage, S., Dellarole, M., Baxter, N.J., Rouget, J.-B., Dimitrov, J.D., Wang, N., Fujimoto, Y., Hounslow, A.M., Lacroix-Desmazes, S., Fukase, K., et al. (2014). Molecular basis for bacterial peptidoglycan recognition by LysM domains. Nat Commun *5*, 4269. https://doi.org/10.1038/ncomms5269.

40. Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology *157*, 105–132. https://doi.org/10.1016/0022-2836(82)90515-0.

41. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine *17*, 405–424. https://doi.org/10.1038/gim.2015.30.

42. Gordon, C., Lu, A.X., and Abbeel, P. (2024). Protein Language Model Fitness Is a Matter of Preference. Preprint, https://doi.org/10.1101/2024.10.03.616542 https://doi.org/10.1101/2024.10.03.616542.

43. Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: A new style of protein science. Nature Methods *11*. https://doi.org/10.1038/nmeth.3027.

44. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*. https://doi.org/10.1038/s41586-020-2308-7.

45. Unsal, S., Atas, H., Albayrak, M., Turhan, K., Acar, A.C., and Doğan, T. (2022). Learning functional properties of proteins with language models. Nat Mach Intell *4*, 227–245. https://doi.org/10.1038/s42256-022-00457-9.

46. Douville, C., Masica, D.L., Stenson, P.D., Cooper, D.N., Gygax, D.M., Kim, R., Ryan, M., and Karchin, R. (2016). Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). Human Mutation *37*, 28–35. https://doi.org/10.1002/humu.22911.

47. Danneels, B., Pinto-Carbó, M., and Carlier, A. (2018). Patterns of Nucleotide Deletion and Insertion Inferred from Bacterial Pseudogenes. Genome Biology and Evolution *10*, 1792–1802. https://doi.org/10.1093/gbe/evy140.

48. Lin, M., Whitmire, S., Chen, J., Farrel, A., Shi, X., and Guo, J. (2017). Effects of short indels on protein structure and function in human genomes. Sci Rep *7*, 9313. https://doi.org/10.1038/s41598-017-09287-x.

49. Taylor, M.S., Ponting, C.P., and Copley, R.R. (2004). Occurrence and Consequences of Coding Sequence Insertions and Deletions in Mammalian Genomes. Genome Res. *14*, 555–566. https://doi.org/10.1101/gr.1977804.

50. Arpino, J.A.J., Reddington, S.C., Halliwell, L.M., Rizkallah, P.J., and Jones, D.D. (2014). Random Single Amino Acid Deletion Sampling Unveils Structural Tolerance and the Benefits of Helical Registry Shift on GFP Folding and Structure. Structure *22*, 889–898. https://doi.org/10.1016/j.str.2014.03.014.

51. Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K.A., Lin, G.N., Nam, H.-J., Mort, M., Cooper, D.N., Sebat, J., Iakoucheva, L.M., et al. (2020). Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun *11*, 5918. https://doi.org/10.1038/s41467-020-19669-x.

52. Shrake, A., and Rupley, J.A. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. Journal of Molecular Biology *79*, 351–371. https://doi.org/10.1016/0022-2836(73)90011-9.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Grant Goldman
6FCF84BB4C974AC...
Author Signature

12/04/2025 | 8:28 AM PST
Date