Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Improved prediction of folding free energy changes with structure-aware and inverse-folding representations

**Prawin Rimal**
Clemson University
**Shailesh Kumar Panday**
Clemson University
**Emil Alexov**
ealexov@clemson.edu

Clemson University

**Article**

**Keywords:** Protein stability, Folding free energy, Mutation effects, Protein language models, Inverse folding, Structure-aware representations

**Posted Date:** November 23rd, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11575271/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Improved prediction of folding free energy changes with structure-aware and inverse-folding representations

Prawin Rimal[1], Shailesh Kumar Panday[1], Emil Alexov[1,2,3*]

[1]Department of Physics and Astronomy, Clemson University, Clemson, 29634, SC, USA
[2]Medical Biophysics Graduate Program, Clemson University, 118 Kinard Laboratory, Clemson, 29634, SC, USA
[3]Clemson University Institute for Human Genetics, 106 Gregor Mendel Circle, Greenwood, 29646, SC, USA

*Corresponding Author: Emil Alexov; ealexov@clemson.edu

ORCID:
Prawin Rimal (P.R.): https://orcid.org/0009-0009-6628-5890
Shailesh Kumar Panday (S.K.P.): https://orcid.org/0000-0003-3099-8679
Emil Alexov (E.A.): https://orcid.org/0000-0001-5346-0156

## Abstract

Thermodynamic stability governs protein folding and biological function, and mutation-induced changes in folding free energy (ΔΔG) can cause loss of function, aggregation, and disease. Predicting these changes across complex mutational landscapes, however, remains challenging. We present SAAFEC-STAIR, which integrates SaProt structure-aware representations with ProteinMPNN inverse-folding information through gated fusion and a supervised mutation-aware prediction head. The model combines residue-level features with whole-protein context, local structural neighborhoods, and substitution-specific information. A tiered evaluation across diverse pretrained protein models separated zero-shot mutation ranking, supervised ΔΔG prediction, and ProteinMPNN fusion. On average, supervised calibration accounted for most of the improvement, with a further gain from ProteinMPNN fusion. Curated, homology-controlled training, validation, and test datasets minimized redundancy and allowed for controlled comparisons with existing methods. On the single-mutation test set, SAAFEC-STAIR achieved Pearson and Spearman correlations of 0.70 and 0.74, respectively, outperforming the evaluated predictors. For double mutations, a pair-corrected extension achieved correlations of 0.60 and 0.62, improving upon the additive single-mutation baseline and the evaluated comparators. These results show how structure-aware and inverse-folding representations can be combined to improve stability prediction across single and double mutations. A freely accessible web server supports prediction and interactive visualization of mutation effects.

**Keywords:** Protein stability; Folding free energy; Mutation effects; Protein language models; Inverse folding; Structure-aware representations

## Introduction

Protein-folding stability is a central biophysical constraint linking amino acid sequence to molecular function [1, 2]. A missense mutation can stabilize or destabilize the folded state, alter folding or unfolding kinetics, promote degradation or aggregation, perturb binding and active sites, or leave global stability nearly unchanged while disrupting a local functional mechanism [3, 4, 5]. The folding free energy change ΔΔG quantifies a major thermodynamic component of how substitutions affect protein abundance, activity, disease risk, and tolerance to sequence variation, making it a central quantity for interpreting and predicting the functional consequences of mutations [6, 7].

Experimental measurement of ΔΔG remains the gold standard, but conventional single-variant workflows are resource-intensive: each variant may require DNA construct preparation through cloning or mutagenesis, protein expression and purification, quality-control checks, replicate biophysical measurements, and thermodynamic model fitting [8, 9]. Differential scanning calorimetry, chemical denaturation, circular dichroism, fluorescence spectroscopy, thermal-shift assays, and proteolysis-based readouts can each provide stability information, but they differ in throughput, directness, underlying assumptions, and potential sources of error [8, 9, 10, 11, 12]. MegaScale cDNA-display proteolysis, a specialized high-throughput assay, has yielded hundreds of thousands of high-quality stability measurements across diverse small protein domains [13]. Nevertheless, these approaches still require specialized library preparation, sequencing, calibration, quality filtering, and interpretation [13].

This experimental complexity motivates the development of computational stability predictors to reduce costs and enable genome-scale investigations. Even a complete single-mutant experimental scan can require thousands of measurements, while the number of measurements for double and higher-order mutants grows combinatorially. Exhaustive experimental ΔΔG measurements are therefore impractical for most proteins and assay conditions. Computational ΔΔG predictors use several broad classes of methods. Physics-based and statistical-potential methods use empirical energy functions, conformational sampling, or knowledge-based potentials [14, 15, 16]; feature-based predictors combine sequence, evolutionary, structural, and biophysical descriptors with regression or other machine-learning models [17, 18, 19]; and deep-learning methods draw on neural networks, geometric learning, protein language models, inverse-folding models, and, increasingly, high-throughput training data [20, 21, 22, 23, 24, 25].

FoldX [14], KORPM [15], and DDGun/DDGun3D [16], for example, estimate mutation effects from empirical energy terms, residue-pair statistics, sequence profiles, or local structural environments and do not depend on large experimental ΔΔG training sets. Feature-based predictors such as PremPS [17], MAESTRO [18], and SAAFEC-SEQ [19] combine structural, evolutionary, physicochemical, or sequence-derived descriptors with regression, random forests, gradient boosting, support vector machines, or consensus learning. These methods were developed largely using ProTherm-derived datasets, including S2648 and related benchmark collections [17, 18, 19, 26]. Neural and representation-learning methods include neural models trained to reproduce physics-based calculations, such as RaSP [22]; structure-aware antisymmetric models, such as ACDC-NN [24]; sequence-based multiscale models, such as MuFaDDG [25]; and geometric three-dimensional structure encoders such as GeoDDG-3D [27]. We previously developed SAAFEC, a structure-based hybrid combining knowledge-modified MM/PBSA and biophysical and statistical terms [28]. SAAFEC-SEQ subsequently extended this approach to sequence-only prediction using physicochemical, evolutionary, PSSM/PsePSSM, and local conservation features, with gradient-boosting models [19] trained on the relatively small S2648 ProTherm-derived single-mutation dataset [26].

Recently, two developments broadened the data and modeling options available for ΔΔG predictions. MegaScale cDNA-display proteolysis provided large datasets of experimental stability measurements for supervised model development, including dense single-mutant and selected multiple-mutant measurements [13]. In parallel, self-supervised protein models enabled the reuse of sequence-only, structure-aware, and multimodal representations learned from large unlabeled corpora [29, 30, 31, 32]. These advances are complementary: MegaScale extends the experimental supervision, while protein language models and inverse-folding networks provide reusable sequence and structural information [13, 29, 30, 33].

Here, we introduce Substituted Amino Acid-Induced Folding free Energy Changes using Structure-aware Transformer and Inverse-folding Representations (SAAFEC-STAIR), a prediction system that estimates ΔΔG for single and double mutations. SAAFEC-STAIR replaces the PSSM-based sequence features used in SAAFEC-SEQ [19] with SaProt structure-aware representations [31] and ProteinMPNN inverse-folding information [33], and is trained on a much larger dataset centered on MegaScale [13] while retaining independent non-MegaScale databases for testing. Removal of redundant entries from the training and testing datasets supports controlled benchmarking of SAAFEC-STAIR against existing methods.

We used a tiered evaluation protocol to address four modeling questions: how much mutation-ranking signal is available directly from frozen pretrained models; how much supervised ΔΔG calibration improves on those frozen representations; whether ProteinMPNN [33] adds useful sequence- and backbone-conditioned information beyond sequence-only, structure-aware, and multimodal protein models; and whether a double-mutation pair-correction head improves on an additive sum of two single-mutant predictions. To answer these questions, we constructed training, validation, and test datasets with duplicate and homology controls, evaluated a broad set of protein-model families using a common protocol, and benchmarked the resulting predictor against existing methods in matched evaluation settings. The final SAAFEC-STAIR configuration was selected based on predictive performance, pretrained model size, and the training data and history of the SaProt checkpoints. The resulting predictor is available through a freely accessible web server.

## Results

### SAAFEC-STAIR integrates SaProt and ProteinMPNN for mutation-stability prediction

Rather than selecting a single representation a priori, we evaluated pretrained protein models across multiple families and input configurations. The tiered protocol compared architecture-specific mutation ranking, supervised ΔΔG prediction from frozen representations, and fusion with ProteinMPNN outputs. These comparisons informed the selection of a three-seed ensemble that combines SaProt 650M (PDB) [31] and ProteinMPNN [33] via gated cross-attention as the final single-mutation predictor.

The selected predictor integrates structure-aware and inverse-folding information (Fig. 1a). SaProt encodes each residue using its amino acid and a Foldseek-derived 3Di token [31, 34], which summarizes tertiary-interaction geometry relative to a spatially nearby residue [34]. ProteinMPNN [33] supplies backbone-conditioned preference scores for the 20 amino acids at each position. Gated cross-attention combines these complementary signals, and the mutation-aware head uses the fused mutation-site state, along with whole-protein context, the local structural neighborhood, mutation identity, and substitution-specific ProteinMPNN features, to predict ΔΔG.

For double mutations, SAAFEC-STAIR reuses this frozen single-mutation ensemble as an additive prior and trains a separate pair-correction head to learn calibration and contact-gated interaction corrections (Fig. 1b). To keep evaluation independent of model development, we used training, validation, and test sets with duplicate and homology controls.

### Curated datasets support independent evaluation

For the single-mutation analysis, we used the MegaScale [13] training, validation, and test partitions together with eight additional benchmark datasets: FireProt(HF) and FireProt [20, 35], S2648 [26, 36], S461 [15], S669 [37], S783 [27], S8754 [27], and Ssym-direct [38]. Their source-specific sizes are reported in Supplementary Table S1. Double-mutation measurements were drawn from MegaScale/cDNA. The complete single- and double-mutation curation workflows are shown in Fig. 2, with data resources provided in Supplementary Table S2 for single mutations and Supplementary Section S1.2 for double mutations.

The MegaScale training partition comprised 239 proteins and 215,731 mutations. To maintain training–validation independence, one homologous validation protein was removed, leaving 30 proteins and 26,554 mutations in the MegaScale validation set; the test set remained unchanged at 28 proteins and 28,172 mutations.

Across these nine single-mutation sources, 47,325 mutation measurements were available before duplicate filtering. After filtering, 35,380 mutations from 412 proteins remained. Consistent duplicate measurements were averaged, and unresolved groups were excluded. Supplementary Fig. S1 shows the protein-mutation overlap among the source databases before duplicate filtering (see Methods).

The subsequent MMseqs2-based homology screen [39] removed 25 proteins and 1,143 mutations, all of which were from non-MegaScale benchmark databases. These proteins and mutations were added to the validation set and were not used for held-out testing. Adding these 25 proteins to the 30-protein MegaScale validation set produced a final validation set of 55 proteins and 27,697 mutations. A final validation–test homology check excluded one additional mutation. The final test set, therefore, contained 386 proteins and 34,236 mutations; Supplementary Section S1.1 and Supplementary Table S3 describe the filtering steps and report the counts.

For the double-mutation dataset, the ThermoMPNN [20] partition and corresponding measurement counts are reported in Supplementary Table S4.

**Tiered evaluation identifies model contributions**

The evaluation protocol comprised three tiers: Tier 0 assessed whether a frozen pretrained model could rank mutations using its own amino acid preference or likelihood outputs; Tier 1 assessed supervised ΔΔG prediction from fixed residue-level embeddings produced by the pretrained model; and Tier 2 assessed whether adding frozen ProteinMPNN logits [33] provided complementary sequence- and backbone-conditioned information (Supplementary Table S5 and Fig. 3). The evaluated protein language models and related pretrained protein models spanned sequence-only, structure-aware, multimodal, and inverse-folding architectures [29, 30, 31, 32, 33, 40, 41, 42, 43, 44, 45, 46, 47, 48]; model configurations, inference inputs, and scoring procedures are detailed in Supplementary Section S4.1 and Supplementary Tables S10 and S11. Tier 0 quantities are architecture-specific relative-preference measures, not ΔΔG predictions or values on a common physical scale; therefore, only Spearman rank correlation was used for Tier 0.

Across the 20 configurations evaluated in all three tiers, performance improved at each stage of the comparison (Fig. 4; Supplementary Table S6). Mean Spearman correlation increased from approximately 0.27 in Tier 0 to 0.64 in Tier 1, an average gain of about 0.37, and rose further to 0.71 after ProteinMPNN fusion in Tier 2. Mean Pearson correlation likewise increased from approximately 0.61 in Tier 1 to 0.67 in Tier 2. Supervised mutation-aware calibration, therefore, produced the largest performance gain, while Tier 2 yielded further improvements in both Spearman and Pearson correlations across all 20 evaluated configurations.

After ProteinMPNN fusion, the largest increases in Spearman correlation occurred for Tranception (0.20), ProGen2 (0.17), and ProteinBERT (0.16), the three models with the lowest Tier 1 correlations (Fig. 4a). In contrast, the increase was 0.01 for SaProt 650M (PDB), 0.02 for ESMC

6B and ESM3-open with sequence and structure, and 0.03 for the ESM3-open full-multimodal configuration. These results show that ProteinMPNN provided complementary information across the evaluated models, although its contribution was smaller for the models with the highest Tier 1 correlations.

SaProt models were among the strongest at both supervised tiers. SaProt 650M (PDB) led Tier 1 with a Spearman correlation of 0.73. After ProteinMPNN fusion, SaProt 650M (PDB), SaProt 650M (AF2), SaProt 1.3B, and ESMC 6B shared the highest reported Tier 2 Spearman and Pearson correlations, 0.74 and 0.70, respectively. Tier 0 performance nevertheless depended on the information and scoring procedure used for each model. Removing the structure-derived 3Di tokens from SaProt 650M (PDB) reduced the Spearman correlation from 0.44 to 0.15; the corresponding correlations were 0.73 versus 0.61 in Tier 1 and 0.74 versus 0.69 in Tier 2, showing that supervised calibration and ProteinMPNN fusion recovered part, but not all, of the lost performance. For PoET-2, the Spearman correlation was 0.08 with the causal score using the query sequence only, 0.18 with masked-residue scoring using the same sequence, and 0.37 with causal scoring additionally supplied with homologous sequences (Supplementary Table S7). Because the homolog-context condition supplied additional evolutionary information, it was reported separately from the primary cross-model comparison. Together, these analyses show that Tier 0 values reflect model-specific scoring and inference context as well as the representations learned during pretraining.

Model size alone did not determine performance among the evaluated models. Among configurations that use only amino acid sequences for inference, ESMC 6B achieved the highest Tier 1 Spearman and Pearson correlations, 0.72 and 0.68, respectively (Fig. 4a and Fig. 4b; Supplementary Table S6). The structure-aware SaProt 650M (PDB) configuration achieved slightly higher correlations of 0.73 and 0.69 despite having substantially fewer parameters. Larger variants of ProtT5, Ankh, and SaProt did not consistently outperform their smaller counterparts. We selected SaProt 650M (PDB) fused with ProteinMPNN as the final single-mutation model, combining strong predictive performance with a preference for fewer pretrained-model parameters. Its 650-million-parameter backbone is smaller than SaProt 1.3B and ESMC 6B, while the corresponding Tier 2 configurations achieved the same reported correlations. Among the two 650M SaProt checkpoints, the PDB version includes additional training on experimentally determined PDB structures and achieved a slightly higher Tier 1 Spearman correlation than the AF2 version (0.73 versus 0.72), with the same reported Tier 1 Pearson and Tier 2 correlations.

**SAAFEC-STAIR improves single-mutation prediction**

Single-mutation performance was benchmarked against a set of existing methods [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 49, 50, 51]; predictor descriptions and software availability are summarized in Supplementary Table S13. The selected three-seed SAAFEC-STAIR single-mutation ensemble achieved Pearson and Spearman correlations of 0.70 and 0.74, respectively, on the final test set (Fig. 5a; Table 1). These were the highest correlations among the evaluated predictors. ThermoMPNN [20] achieved Pearson and Spearman correlations of 0.67 and 0.70, respectively, whereas SPURS [21], the closest comparator, achieved 0.68 and 0.71. GeoDDG-3D [27], MutateEverything [49], and PROSTATA [50] required method-specific homology control

and were evaluated on separate filtered subsets of the final test set (Fig. 5b; Supplementary Table S8).

On these filtered subsets, SAAFEC-STAIR outperformed all three comparators (Fig. 5b; Supplementary Table S8). Because each filter retained a different set of mutations, comparisons should be interpreted within each matched subset.

### Pair correction improves double-mutation prediction

We next asked whether a learned pair correction could improve the additive estimate while leaving the three-seed single-mutation ensemble frozen. A separate trainable head estimates calibration and contact-gated interaction corrections from symmetric pair context (Fig. 1b).

Pair-corrected SAAFEC-STAIR was benchmarked against several existing methods supporting double-mutation prediction [14, 21, 51, 52]. On the complete double-mutation test set, pair-corrected SAAFEC-STAIR achieved Pearson and Spearman correlations of 0.60 and 0.62, respectively. The additive single-mutation baseline reached 0.56 and 0.60. Thus, the learned pair correction increased Pearson correlation by 0.04 and Spearman correlation by 0.02, although the gains were modest. On the complete test set, SAAFEC-STAIR achieved the highest Pearson and Spearman correlations, followed by SPURS [21] and ThermoMPNN-D [52] (Fig. 6a; Table 2).

For the separate overlap-controlled comparison, 13 of the 20 complete-test proteins had identifiers that appeared in the published cDNA training split and were removed, leaving 4,599 double mutations from seven proteins. MutateEverything [49] was evaluated on this seven-protein subset (Fig. 6b; Supplementary Table S9). On this subset, SAAFEC-STAIR achieved a Pearson correlation of 0.49, compared with 0.46 for MutateEverything; both methods had a Spearman correlation of 0.40 after rounding. The comparison is therefore limited to these seven proteins.

## Discussion

The tiered analysis showed that supervised mutation-aware calibration accounted for most of the improvement beyond architecture-specific Tier 0 mutation ranking, while ProteinMPNN fusion provided a smaller additional gain. Variation among the architecture-specific Tier 0 scores further shows that zero-shot performance depends on the pretraining objective, scoring formulation, and information supplied during inference. Tier 0 should therefore be interpreted as model-specific zero-shot ranking rather than as a uniform measure of representation quality; a weak Tier 0 ranking does not imply that a model lacks useful representations for supervised prediction.

Training the common mutation-aware prediction head on experimental ΔΔG values reduced variation in correlation coefficients across model configurations in Tier 1, and adding ProteinMPNN in Tier 2 further reduced this variation. Consistent with this pattern, stronger Tier 1 configurations, including SaProt [31] and the ESM3-open configurations with structural input [32], showed smaller gains from ProteinMPNN fusion [33], whereas several models with lower Tier 1 performance showed larger gains. In the SaProt ablation, ProteinMPNN recovered some, but not all, of the performance lost when 3Di tokens were removed. Together, these results suggest that the common mutation-aware head extracts signals shared across pretrained representations, and that SaProt and ProteinMPNN provide partially complementary information. However, the performance differences do not indicate distinct physical contributions from the two inputs.

Although the Tier 2 Pearson and Spearman correlations were relatively close across models, their remaining differences indicate that the original pretrained representations continued to influence performance after fusion. The narrow range of correlations likely reflects the present dataset and modeling framework rather than a general performance ceiling.

Tier 1 provides the clearest basis for comparing the pretrained protein models because it applies the common mutation-aware prediction framework without the additional ProteinMPNN stream. The models nevertheless retain their respective input modalities and representation schemes. Larger variants of ProtT5 [30] and SaProt [31] did not consistently outperform their smaller counterparts, whereas scaling improved performance within ESM-2 and ESMC [40]; the larger Ankh variants [41, 42] likewise showed no consistent advantage. Therefore, input modality, the task used during pretraining, training data, and prediction setup appear to matter alongside parameter count.

The matched input comparisons also show that the contribution of structural information differs among models. Adding structural input to ESM3-open [32] produced a modest improvement in Tiers 0 and 1, while the complete multimodal input provided no further improvement at the reported precision. In contrast, removing 3Di tokens from SaProt led to a greater reduction across both tiers. The strong Tier 1 performance of the sequence-only ESMC 6B model shows that explicit structural input is neither necessary nor sufficient for high predictive performance. ESM3-open produced stronger zero-shot rankings than the structure-aware SaProt configurations, whereas the ordering reversed in Tier 1, illustrating that native zero-shot ranking and performance after supervised calibration are related but distinct properties. SaProt 650M (PDB) achieved one of the highest Tier 1 correlations, using substantially fewer parameters than several other evaluated models.

SPURS also combines a protein language model with ProteinMPNN to predict mutation stability, but its fusion strategy differs from that of SAAFEC-STAIR. SPURS rewires ProteinMPNN structural representations into ESM-2 [21], whereas SAAFEC-STAIR retains separate SaProt and ProteinMPNN streams and combines ProteinMPNN output logits within the downstream prediction head. The two methods, therefore, use related sources of structural information in different computational roles, and their performance should be interpreted for each prediction system as a whole rather than attributed solely to the presence of ProteinMPNN.

Similarity between test proteins and proteins used during model development can inflate apparent performance on new proteins. Accordingly, method-specific homology-filtered subsets were used for predictors with substantial overlap between their training proteins and the final test set. These controls reduce a potential source of bias, although the resulting evaluation subsets differ among the affected predictors.

For double mutations, the pair-corrected model improved both Pearson and Spearman correlations relative to the additive single-mutation baseline on the complete double-mutation test set. The larger improvement in Pearson correlation than in Spearman correlation is consistent with the pair-specific terms refining predicted magnitudes more strongly than they reorder mutations across the complete dataset. Because the interaction term was trained to improve prediction, it should not be interpreted as a direct measure of epistasis or as evidence of uniform improvement across proteins [52].

Although the final test set included proteins from several sources, the MegaScale component consists mainly of small domains [13]. Performance on larger multidomain and membrane proteins is therefore less well established. The curated ΔΔG databases differ in assay conditions and data sources [8, 9, 26, 35], and our analysis cannot determine how much experimental heterogeneity contributes to variation in performance. The model uses wild-type structures without relaxing mutant structures, and local geometric errors in predicted structures may affect the structure-dependent representations.

## Methods

### Free-energy convention and dataset construction

Mutation-induced folding free-energy changes were defined as

$$\Delta\Delta G = \Delta G_{\mathrm{mut}} - \Delta G_{\mathrm{WT}} \quad (1)$$

where $\Delta G_{mut}$ and $\Delta G_{WT}$ denote the folding free energies of the mutant and wild-type proteins, respectively. Under this convention, $\Delta\Delta G > 0$ indicates destabilization and $\Delta\Delta G < 0$ indicates stabilization; values are reported in kcal/mol [8, 9].

For the single-mutation analysis, the principal high-throughput source was MegaScale, which estimates folding stability from cDNA-display proteolysis using trypsin- and chymotrypsin-survival curves, K50-related measurements, and a kinetic/thermodynamic model [13]. We used the protein-level MegaScale train/validation/test partition established by ThermoMPNN [20]. The single-mutation test compilation additionally included FireProt(HF) and FireProt [20, 35], S2648 [26, 36], S461 [15], Ssym-direct [38], S669 [37], S783 [27], and S8754 [27]. The source databases and their sizes before duplicate filtering are summarized in Supplementary Table S1. These databases overlap at the protein-mutation level: FireProt(HF) is contained within FireProt, S461 is contained within S669, and S2648 is nearly contained within S8754 under this matching definition. These overlaps were addressed during dataset construction and filtering.

For held-out evaluation, the MegaScale proteins were partitioned after clustering with MMseqs2 using a 25% sequence-identity cutoff. The MegaScale validation proteins were checked for homology with the training proteins before compiling the final test set. We constructed the final test set in two stages: first, measurements duplicated across databases were resolved for each protein-mutation pair; second, the remaining test proteins were screened against the MegaScale training and validation proteins to remove sequence homologs.

Single measurements were kept as reported. When two measurements were available, their ΔΔG values were averaged only if their signs agreed and their difference did not exceed 1.0 kcal/mol. For groups containing at least three measurements, averaging required consistent signs and a sample standard deviation of at most 1.0 kcal/mol. One outlier could be removed when doing so allowed a larger group to satisfy these conditions; duplicate groups that still failed were excluded.

In the second stage, MMseqs2 was used to compare the duplicate-filtered test proteins against the 239 training proteins and 30 MegaScale validation proteins. Proteins matching either set at ≥25% sequence identity and ≥80% coverage were excluded from held-out testing, and the excluded proteins and mutations were incorporated into the validation set. A final homology check compared the remaining test proteins with the expanded validation set (Supplementary Table S3).

For the double-mutation analysis, measurements were drawn from MegaScale/cDNA and assigned using the protein-level training, validation, and test partition established by ThermoMPNN. In that partition, MegaScale proteins were clustered using MMseqs2 with a minimum identity of 25%, and entire clusters were assigned to one of the three splits.

**Tiered evaluation of pretrained models**

Candidate pretrained models were evaluated under the three-tier protocol summarized in Fig. 3 and detailed below. For this comparison, we evaluated protein language models (PLMs) and related pretrained protein models spanning several architectures and input modalities: the bidirectional sequence encoders ESM-2 [29] and ESMC [40]; the encoder–decoder models ProtT5 [30, 43], Ankh Large [41], and Ankh3 XL [42], for which the Tier 1 and Tier 2 prediction heads received encoder-derived residue representations; the autoregressive models ProGen2 [44] and Tranception [45]; the convolutional CARP model [46]; PoET-2 [47] and ProteinBERT [48]; the structure-aware SaProt family [31]; and ESM3 under sequence-only, sequence-plus-structure, and multimodal conditions [32]. ProteinMPNN [33] was analyzed separately from the PLMs as a backbone-conditioned inverse-folding model predicting amino acid preferences at each residue. ProteinMPNN was evaluated as a standalone baseline in Tiers 0 and 1, and its residue-aligned logits supplied the additional backbone-conditioned information used in Tier 2. The corresponding software repositories and model-weight sources are listed in Supplementary Table S11.

Across these models, Tier 0 scoring was tailored to each pretrained protein model's architecture, whereas Tiers 1 and 2 used a common downstream supervised framework. The input modality of each pretrained-model branch was retained across tiers: sequence-only models used amino acid sequences, whereas multimodal models also used the structural tracks specified for that configuration. For models requiring structural input, ColabFold was used to generate a wild-type structure for each protein [53, 54]; separate mutant structures were not generated. Tier 2 additionally received the ProteinMPNN stream; therefore, applicable Tier 2 versions of otherwise sequence-only models also incorporated backbone-derived information. Accordingly, within the same pretrained model and input conditions, differences between tiers reflected the addition of supervised calibration and, subsequently, the ProteinMPNN stream. Comparisons across model families remained descriptive because model size, architecture, and pretraining data differed simultaneously.

**Tier 0 (zero-shot mutation ranking).** For a proposed substitution at residue $i$, $w_i$ denotes the wild-type amino acid and $a$ denotes the candidate mutant amino acid. Tier 0 was evaluated as a zero-shot ranking task without ΔΔG supervision: the pretrained model remained frozen, no prediction head was fitted, and experimental ΔΔG measurements were not used to construct the mutation scores. Instead, the frozen model produced a relative preference score comparing the candidate mutant with the wild type. When the model provided amino acid probabilities at the mutation site, the destabilization-oriented score was calculated as ln P($w_i$ | context) − ln P($a$ | context). A candidate assigned a lower probability than the wild type therefore received a positive value and was ranked as more destabilizing, whereas a candidate favored over the wild type received a negative value. This relative preference is not a calibrated ΔΔG estimate and does not have a kcal/mol interpretation. Masked, causal, autoregressive, and encoder–decoder models

yield different probability or preference outputs; Tier 0, therefore, used a scoring procedure aligned with each model's pretraining architecture. When an optional homolog context was available, the primary cross-model comparison used query-sequence-only inputs to preserve comparability. Model-specific procedures are provided in Supplementary Section S4.1 and Supplementary Table S10. Tier 0 was evaluated using Spearman correlation with experimental ΔΔG.

**Tier 1 (supervised mutation-aware prediction).** Each pretrained protein model remained frozen, while a common mutation-aware prediction head was trained on experimental ΔΔG measurements. The head projected residue embeddings into a shared hidden space and combined mutation-site and whole-protein context with substitution identity to produce a wild-type-centered L × 20 matrix. Only experimentally measured substitutions were used for training or evaluation. Each model retained its assigned input modality, and an additional ProteinMPNN stream was introduced only in Tier 2. The selected model's prediction head is specified below.

**Tier 2 (ProteinMPNN-augmented fusion).** This tier used frozen pretrained representations and the Tier 1 mutation-aware head architecture while adding residue-aligned ProteinMPNN logits generated from the wild-type backbone and sequence using the ThermoMPNN implementation [20] and official ProteinMPNN checkpoint [33]. Because this stream was backbone-conditioned, all applicable Tier 2 configurations included structural information. Gated cross-attention [55] fused the two streams before the common wild-type-centered readout. The selected SaProt–ProteinMPNN fusion architecture and substitution-specific ProteinMPNN features are specified below and in Supplementary Section S4.2.1.

### SaProt–ProteinMPNN gated fusion architecture

For a wild-type protein of length L, let S denote the matrix of frozen SaProt residue embeddings, Z the matrix of frozen ProteinMPNN logits, and R the matrix of wild-type Cα coordinates. Each row of S is a 1,280-dimensional SaProt embedding [31], each row of Z contains 20 ProteinMPNN amino-acid compatibility scores [33], and each row of R contains the three Cartesian coordinates of one Cα atom. Thus,

$$\boldsymbol{S} \in \mathbb{R}^{L\times 1280}, \boldsymbol{Z} \in \mathbb{R}^{L\times 20}, \boldsymbol{R} \in \mathbb{R}^{L\times 3} \tag{2}$$

For residue $i$, $S_i$ and $Z_i$ denote the corresponding rows of S and Z. In the equations below, these rows are treated as column vectors, and *LN* denotes layer normalization. Learned matrices $W^{\mathrm{S}}$ and $W^{\mathrm{M}}$ project the SaProt and ProteinMPNN inputs, respectively, into a shared 768-dimensional space, producing $H_{\mathrm{i}}^{\mathrm{S}}$ and $H_{\mathrm{i}}^{\mathrm{M}}$:

$$\boldsymbol{H}_i^S = \boldsymbol{W}^S \mathrm{LN}(\boldsymbol{S}_i), \boldsymbol{H}_i^M = \boldsymbol{W}^M \mathrm{LN}(\boldsymbol{Z}_i) \tag{3}$$

Both projected states lie in $\mathbb{R}^{768}$. Stacking the residue-level projected states over all positions gives the length-by-768 matrices used as the attention inputs below. During supervised training, only the downstream fusion and prediction head were optimized. To fuse the projected SaProt and ProteinMPNN streams, cross-attention uses the projected SaProt states as the query matrix Q and the projected ProteinMPNN states as the key matrix K and value matrix V [55]:

$$\boldsymbol{Q} = \boldsymbol{H}^S \tag{4a}$$

$$\boldsymbol{K} = \boldsymbol{V} = \boldsymbol{H}^{M} \tag{4b}$$

$$\boldsymbol{A} = \mathrm{MHA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) \tag{4c}$$

Here, *MHA* denotes eight-head multi-head attention, and *A* is the resulting attended ProteinMPNN representation. This operation allows the SaProt representation at each residue to incorporate ProteinMPNN information from positions across the protein. For each residue $i$, the projected SaProt state $H_{\mathrm{i}}^{\mathrm{S}}$, the directly projected ProteinMPNN state $H_{\mathrm{i}}^{\mathrm{M}}$, and the attended ProteinMPNN signal $A_i$ are concatenated:

$$\boldsymbol{C}_i = \left[\boldsymbol{H}_i^{S}; \boldsymbol{H}_i^{M}; \boldsymbol{A}_i\right] \tag{5a}$$

This 2,304-dimensional combined state is passed through a gate-specific multilayer perceptron, followed by a logistic sigmoid to produce the 768-dimensional gate $G_{\mathrm{i}}$:

$$\boldsymbol{G}_i = \sigma\left(\mathrm{MLP}_{\mathrm{gate}}(\boldsymbol{C}_i)\right) \tag{5b}$$

Each component of $G_i$ lies between 0 and 1. Values close to 1 emphasize the cross-attended ProteinMPNN signal $A_i$, whereas values close to 0 emphasize the directly projected ProteinMPNN state $H_{\mathrm{i}}^{\mathrm{M}}$. Because the gate is calculated separately for each residue and hidden feature, it can vary this balance across both positions and feature channels. The gated combination of the two ProteinMPNN signals is then added to the projected SaProt state:

$$\boldsymbol{F}_i = \mathrm{LN}\left(\boldsymbol{H}_i^{S} + \boldsymbol{G}_i \odot \boldsymbol{A}_i + (\boldsymbol{1} - \boldsymbol{G}_i) \odot \boldsymbol{H}_i^{M}\right) \tag{5c}$$

Here, $\odot$ denotes element-wise multiplication, and layer normalization yields the fused residue representation $F_i$.

**Mutation-aware prediction head and ensemble**

For each proposed substitution, the prediction head supplements the fused state at the target position with whole-protein, local-structural, and substitution-specific information. The whole-protein context $P$ is obtained by averaging the fused states over all $L$ residues and applying the learned projection $W^P$:

$$\boldsymbol{P} = \boldsymbol{W}^{P}\left(\frac{1}{L}\sum_{j=1}^{L}\boldsymbol{F}_j\right) \tag{6}$$

The resulting 768-dimensional vector is shared across all residue positions and possible substitutions in the protein. In parallel, the local-neighborhood branch uses wild-type Cα distances to compute a distance-weighted average of the fused states surrounding residue $i$. It excludes the target residue itself, retains residues within 10 Å, and uses at most the 16 neighbors with the largest weights. The resulting neighborhood summary is projected into the shared 768-dimensional space and denoted $N_i$. The distance weights, normalization, and treatment of residues without valid neighbors are described in Supplementary Section S4.2.1, Eqs. (S1a) and (S1b).

For a substitution at residue $i$, $w_i$ denotes the wild-type amino acid and $a$ denotes the proposed mutant amino acid. The mutation-identity block represents both amino-acid identities, the difference between their learned embeddings, and an ordered wild-type-to-mutant pair

embedding. The ordered pair distinguishes substitutions such as A→V and V→A, allowing predictions at the same residue to depend on the direction and identity of the proposed change.

Finally, ProteinMPNN contributes a separate substitution-specific feature block. Let $p_i(x)$ = softmax$(Z_i)_x$ denote the probability assigned to amino acid $x$ at residue $i$. For substitution $w_i \rightarrow a$, this block contains ln $p_i(w_i)$, ln $p_i(a)$, their difference ln $p_i(a)$ - ln $p_i(w_i)$ and the Shannon entropy of the 20-amino-acid distribution. Entropy indicates whether the full distribution is concentrated on a few amino acids or spread more broadly and is defined in Supplementary Eq. (S1c). The mutation-identity and ProteinMPNN blocks are then projected separately to 768 dimensions for combination with $F_i$, $P$, and $N_i$.

To assemble the prediction matrix for substitution $w_i \rightarrow a$, let $H_{i,a}^{\text{sub}}$ denote the projected mutation-identity block and $H_{i,a}^{\text{MPNN}}$ the projected ProteinMPNN feature block. These two substitution-specific vectors, along with the previously defined $F_i$, $P$, and $N_i$, are all 768-dimensional. The model adds the five terms to form the complete mutation-specific representation:

$$\boldsymbol{X}_{i,a} = \boldsymbol{F}_i + \boldsymbol{P} + \boldsymbol{N}_i + \boldsymbol{H}_{i,a}^{\text{sub}} + \boldsymbol{H}_{i,a}^{\text{MPNN}} \tag{7}$$

The complete representation is processed by a trainable score network:

$$R_{i,a} = f_\theta^{\text{score}}\left(\boldsymbol{X}_{i,a}\right) \tag{8a}$$

Here, $f_\theta^{\text{score}}$: $\mathrm{R}^{768} \rightarrow \mathrm{R}$ denotes the trainable prediction head, comprising two residual feed-forward blocks followed by a scalar linear output layer. It maps each mutation-specific representation $X_{\text{i,a}}$ to an unconstrained raw score $R_{i,a}$.

The predicted stability change, denoted $\Delta\Delta G_{\text{i,a}}$, is obtained by subtracting the score assigned to the wild-type amino acid $w_i$ at residue $i$ from $R_{i,a}$. Once the frozen wild-type representations have been computed, the mutation-specific representations are evaluated in parallel. For a protein of length L, a single forward pass through each ensemble member's prediction head yields a complete L × 20 saturation mutagenesis matrix, containing predictions for all 19 non-wild-type substitutions at each residue and the corresponding wild-type (no-change) channel fixed at zero:

$$\Delta\Delta G_{i,a} = R_{i,a} - R_{i,w_i} \tag{8b}$$

Only measured substitutions contribute to training; unmeasured entries are masked. Training, validation, and optimization settings are provided below and in Supplementary Table S12. Let *k = 1, 2, 3* index the three models trained from different random initializations. Their complete prediction matrices are averaged to obtain the final ensemble prediction and reduce variability associated with random initialization:

$$\Delta\Delta G_{i,a}^{\text{ens}} = \frac{1}{3}\sum_{k=1}^{3} \Delta\Delta G_{i,a}{}^{(k)} \tag{9}$$

The workflow can also generate an interactive HTML view of the final ensemble prediction matrix for examining individual residues and substitutions; the interface is described in Supplementary Section S4.2.2 and shown in Supplementary Fig. S2.

**Pair-corrected double-mutation model**

The pair-corrected double-mutation model uses the frozen three-seed single-mutation ensemble and does not retrain the SaProt–ProteinMPNN pathway (Fig. 1b). Let $D$ denote the ensemble single-mutation prediction matrix. For a specified pair of substitutions $w_{\mathrm{i}} \rightarrow a$ at residue $i$ and $w_{\mathrm{j}} \rightarrow b$ at residue $j$, $D_{\mathrm{i,a}}$ and $D_{\mathrm{j,b}}$ are the corresponding single-mutation predictions. Their sum, $B_{\mathrm{ij}}^{\mathrm{ab}}$, is the additive baseline:

$$B_{ij}^{ab} = D_{i,a} + D_{j,b} \tag{10}$$

In Equations (10) and (11), *i* and *j* denote the two residue positions, whereas *a* and *b* denote the corresponding proposed amino acids. For each site, the double-mutation head combines the frozen SaProt residue embedding, the wild-type and mutant amino-acid identities, and the corresponding single-mutation prediction. The two site descriptions are combined symmetrically and supplemented with whole-protein context, summaries of the two single-mutation predictions, wild-type Cα geometry, and ProteinMPNN features at both sites. These features inform the two correction branches.

Within this head, the calibration branch produces the pair-specific term $\Delta_{\mathrm{cal,ij}}^{\mathrm{ab}}$, which adjusts the additive baseline's scale and offset. The interaction branch produces a bounded pair correction. A learned residual gate controls the strength of this correction, while a smooth multiplier based on the wild-type Cα distance permits stronger corrections for nearby sites and shrinks, without eliminating, corrections for distant sites. Their product is the interaction residual $\Delta_{\mathrm{int,ij}}^{\mathrm{ab}}$. The predicted double-mutation stability change $\Delta\Delta G_{\mathrm{ij}}^{\mathrm{ab}}$ is therefore the sum of the additive baseline and these two corrections:

$$\Delta\Delta G_{ij}^{ab} = B_{ij}^{ab} + \Delta_{\mathrm{cal},ij}^{ab} + \Delta_{\mathrm{int},ij}^{ab} \tag{11}$$

Only the approximately 3.86-million-parameter double-mutation head is trained; SaProt, ProteinMPNN, and the single-mutation ensemble remain frozen. The interaction residual represents a statistical departure from additivity and should not be interpreted as a direct measure of epistasis.

**Model training, selection and evaluation**

Training, validation, and test proteins were kept separate: training proteins supplied gradient updates, validation proteins were reserved for model selection, and test proteins were reserved for final evaluation. The supervised objective combined a masked, weighted mean-squared-error term with a within-protein ranking term. Let $\mathcal{L}_{\mathrm{wMSE}}$ denote the weighted mean-squared-error term and $\mathcal{L}_{\mathrm{rank}}$ the within-protein ranking term; the objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{wMSE}} + 0.15\mathcal{L}_{\mathrm{rank}} \tag{12}$$

The coefficient 0.15 controls the contribution of the ranking term. Validation performance was used for early stopping and checkpoint selection. For each independently initialized single-mutation model, the checkpoint with the best validation performance, measured by Pearson correlation, was retained; the three selected checkpoints were combined according to Equation (9) to form the final ensemble. The single-mutation ensemble was then kept frozen while the

double-mutation pair-correction head was trained, and its checkpoint was selected based on the highest global validation Pearson correlation. Optimizer, learning rates, weight decay, and the remaining training and ensemble settings are summarized in Supplementary Table S12.

After model selection, Pearson and Spearman correlations were calculated between experimental measurements and predictor outputs after alignment to the ΔΔG sign convention defined by Equation (1). The dataset- and predictor-specific sign adjustments are detailed in Supplementary Section S4.5.2. Existing predictors were included when standalone code was available and batch-scale prediction could be completed for the benchmark proteins and mutations. Their methodological descriptions, training or parameter sources, and availability are summarized in Supplementary Table S13.

For PROSTATA, the ESM-2 configuration of MutateEverything, and GeoDDG-3D, substantial similarity was detected between proteins represented in their training data and proteins in the final test set. These three predictors were therefore compared with SAAFEC-STAIR only on separate method-specific homology-filtered subsets of that test set (<25% sequence identity and no detected training homology). Within each filtered subset, performance metrics for SAAFEC-STAIR and the corresponding existing predictor were calculated on the same retained mutations. The subset sizes are summarized in Supplementary Table S14. For the double-mutation comparison with MutateEverything, a seven-protein control subset was created by removing every test protein whose identifier appeared in the published cDNA training split. Additional overlap details are provided in Supplementary Section S3.2.

### Use of artificial intelligence

ChatGPT and Grammarly were used to assist with language editing of the manuscript. The authors reviewed and approved the final manuscript and take full responsibility for its content.

### Data availability

The source datasets analyzed in this study are publicly available through the resources listed in Supplementary Table S2 and Supplementary Section S1.2. The training, validation, and test datasets used by SAAFEC-STAIR, together with the Tier 0–2 predictions and existing-predictor evaluation results, are available at https://github.com/cosmicprawin-ux/saafec-stair and https://compbio.clemson.edu/saafec-stair/.

### Code availability

A public web server is available at https://compbio.clemson.edu/saafec-stair/, where users can upload protein structures, specify single or double mutations, obtain predicted ΔΔG values, download prediction files, and access the visualization interface. The standalone SAAFEC-STAIR inference implementation and the selected-model training workflows are available at https://github.com/cosmicprawin-ux/saafec-stair. The trained SAAFEC-STAIR inference checkpoints are also included in the GitHub repository.

### References

1. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181,** 223-230 (1973).

2. Dill, K. A. & MacCallum, J. L. The protein-folding problem, 50 years on. *Science* **338,** 1042-1046 (2012).
3. Dobson, C. M. Protein folding and misfolding. *Nature* **426,** 884-890 (2003).
4. Chiti, F. & Dobson, C. M. Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade. *Annu. Rev. Biochem.* **86,** 27-68 (2017).
5. Yue, P., Li, Z. & Moult, J. Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* **353,** 459-473 (2005).
6. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19,** 596-604 (2009).
7. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl Acad. Sci. USA* **103,** 5869-5874 (2006).
8. Robertson, A. D. & Murphy, K. P. Protein structure and the energetics of protein stability. *Chem. Rev.* **97,** 1251-1268 (1997).
9. Lindorff-Larsen, K. & Teilum, K. Linking thermodynamics and measurements of protein stability. *Protein Eng. Des. Sel.* **34,** gzab002 (2021).
10. Pace, C. N. Determination and analysis of urea and guanidine hydrochloride denaturation curves. *Methods Enzymol.* **131,** 266-280 (1986).
11. Niesen, F. H., Berglund, H. & Vedadi, M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nat. Protoc.* **2,** 2212-2221 (2007).
12. Park, C. & Marqusee, S. Pulse proteolysis: a simple method for quantitative determination of protein stability and ligand binding. *Nat. Methods* **2,** 207-212 (2005).
13. Tsuboyama, K. et al. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620,** 434-444 (2023).
14. Guerois, R., Nielsen, J. E. & Serrano, L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320,** 369-387 (2002).
15. Hernández, I. M., Dehouck, Y., Bastolla, U., López-Blanco, J. R. & Chacón, P. Predicting protein stability changes upon mutation using a simple orientational potential. *Bioinformatics* **39,** btad011 (2023).
16. Montanucci, L., Capriotti, E., Frank, Y., Ben-Tal, N. & Fariselli, P. DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. *BMC Bioinformatics* **20,** 335 (2019).
17. Chen, Y. et al. PremPS: predicting the impact of missense mutations on protein stability. *PLoS Comput. Biol.* **16,** e1008543 (2020).
18. Laimer, J., Hofer, H., Fritz, M., Wegenkittl, S. & Lackner, P. MAESTRO - multi agent stability prediction upon point mutations. *BMC Bioinformatics* **16,** 116 (2015).
19. Li, G., Panday, S. K. & Alexov, E. SAAFEC-SEQ: a sequence-based method for predicting the effect of single point mutations on protein thermodynamic stability. *Int. J. Mol. Sci.* **22,** 606 (2021).
20. Dieckhaus, H., Brocidiacono, M., Randolph, N. Z. & Kuhlman, B. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proc. Natl Acad. Sci. USA* **121,** e2314853121 (2024).
21. Li, Z. & Luo, Y. Generalizable and scalable protein stability prediction with rewired protein generative models. *Nat. Commun.* **17,** 891 (2026).
22. Blaabjerg, L. M. et al. Rapid protein stability prediction using deep learning representations. *eLife* **12,** e82593 (2023).
23. Gong, J. et al. THPLM: a sequence-based deep learning framework for protein stability changes prediction upon point variations using pretrained protein language model. *Bioinformatics* **39,** btad646 (2023).
24. Benevenuta, S., Pancotti, C., Fariselli, P., Birolo, G. & Sanavia, T. An antisymmetric neural network to predict free energy changes in protein variants. *J. Phys. D Appl. Phys.* **54,** 245403 (2021).

25. Gong, J. et al. MuFaDDG: a sequence-based multiscale feature fusion framework for protein stability changes prediction. *Bioinformatics* **42,** btag196 (2026).
26. Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K. & Sarai, A. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.* **32,** D120-D121 (2004).
27. Xu, Y., Liu, D. & Gong, H. Improving the prediction of protein stability changes upon mutations by geometric learning and a pre-training strategy. *Nat. Comput. Sci.* **4,** 840-850 (2024).
28. Getov, I., Petukh, M. & Alexov, E. SAAFEC: predicting the effect of single point mutations on protein folding free energy using a knowledge-modified MM/PBSA approach. *Int. J. Mol. Sci.* **17,** 512 (2016).
29. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379,** 1123-1130 (2023).
30. Elnaggar, A. et al. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44,** 7112-7127 (2022).
31. Su, J. et al. SaProt: protein language modeling with structure-aware vocabulary. In: *The Twelfth International Conference on Learning Representations* (2024).
32. Hayes, T. et al. Simulating 500 million years of evolution with a language model. *Science* **387,** 850-858 (2025).
33. Dauparas, J. et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378,** 49-56 (2022).
34. van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42,** 243-246 (2024).
35. Stourac, J. et al. FireProtDB: database of manually curated protein stability data. *Nucleic Acids Res.* **49,** D319-D324 (2021).
36. Dehouck, Y. et al. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics* **25,** 2537-2543 (2009).
37. Pancotti, C. et al. Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. *Brief. Bioinform.* **23,** bbab555 (2022).
38. Pucci, F., Bernaerts, K. V., Kwasigroch, J. M. & Rooman, M. Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics* **34,** 3659-3665 (2018).
39. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35,** 1026-1028 (2017).
40. Candido, S. et al. Language modeling materializes a world model of protein biology. *bioRxiv*, https://doi.org/10.64898/2026.06.03.729735 (2026).
41. Elnaggar, A. et al. Ankh: optimized protein language model unlocks general-purpose modelling. *arXiv*, https://doi.org/10.48550/arXiv.2301.06568 (2023).
42. Alsamkary, H., Elshaffei, M., Elkerdawy, M. & Elnaggar, A. Ankh3: multi-task pretraining with sequence denoising and completion enhances protein representations. *arXiv*, https://doi.org/10.48550/arXiv.2505.20052 (2025).
43. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21,** 1-67 (2020).
44. Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N. & Madani, A. ProGen2: exploring the boundaries of protein language models. *Cell Syst.* **14,** 968-978.e3 (2023).
45. Notin, P. et al. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In: *Proceedings of the 39th International Conference on Machine Learning* **162,** 16990-17017. PMLR (2022).

46. Yang, K. K., Fusi, N. & Lu, A. X. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Syst.* **15,** 286-294.e2 (2024).
47. Truong, T. F., Jr. & Bepler, T. Understanding protein function with a multimodal retrieval-augmented foundation model. In: *Advances in Neural Information Processing Systems* **38,** 53167-53222 (2025).
48. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38,** 2102-2110 (2022).
49. Ouyang-Zhang, J., Diaz, D. J., Klivans, A. R. & Krähenbühl, P. Predicting a protein's stability under a million mutations. In: *Advances in Neural Information Processing Systems* **36,** 76229-76247. Neural Information Processing Systems Foundation, Inc. (NeurIPS) (2023).
50. Umerenkov, D. et al. PROSTATA: a framework for protein stability assessment using transformers. *Bioinformatics* **39,** btad671 (2023).
51. Barducci, G. et al. JanusDDG: a physics-informed neural network for sequence-based protein stability via two-fronts attention. *Commun. Biol.* **9,** 494 (2026).
52. Dieckhaus, H. & Kuhlman, B. Protein stability models fail to capture epistatic interactions of double point mutations. *Protein Sci.* **34,** e70003 (2025).
53. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19,** 679-682 (2022).
54. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583-589 (2021).
55. Vaswani, A. et al. Attention is all you need. In: *Advances in Neural Information Processing Systems* **30,** 5998-6008. Curran Associates, Inc. (2017).

## Acknowledgements

This work was funded by the NIH grant number R35GM151964. The computational work reported in this study was performed using the Palmetto Cluster, Clemson University's high-performance computing resource.

## Author contributions

P.R. and E.A. conceived and designed the study. P.R. curated and processed the benchmark datasets, developed the machine-learning methodology and model architectures, performed the computational experiments, and analyzed and visualized the results. P.R. and S.K.P. developed the standalone software package, and S.K.P. developed the web-server interface. P.R. wrote the original draft of the manuscript. P.R., S.K.P. and E.A. reviewed and revised the manuscript. E.A. supervised the study, provided computational resources, and acquired funding. All authors read and approved the final manuscript.

## Additional Information

## Competing interests

The authors declare no competing interests.

Supplementary Information accompanies this manuscript.

## Figures

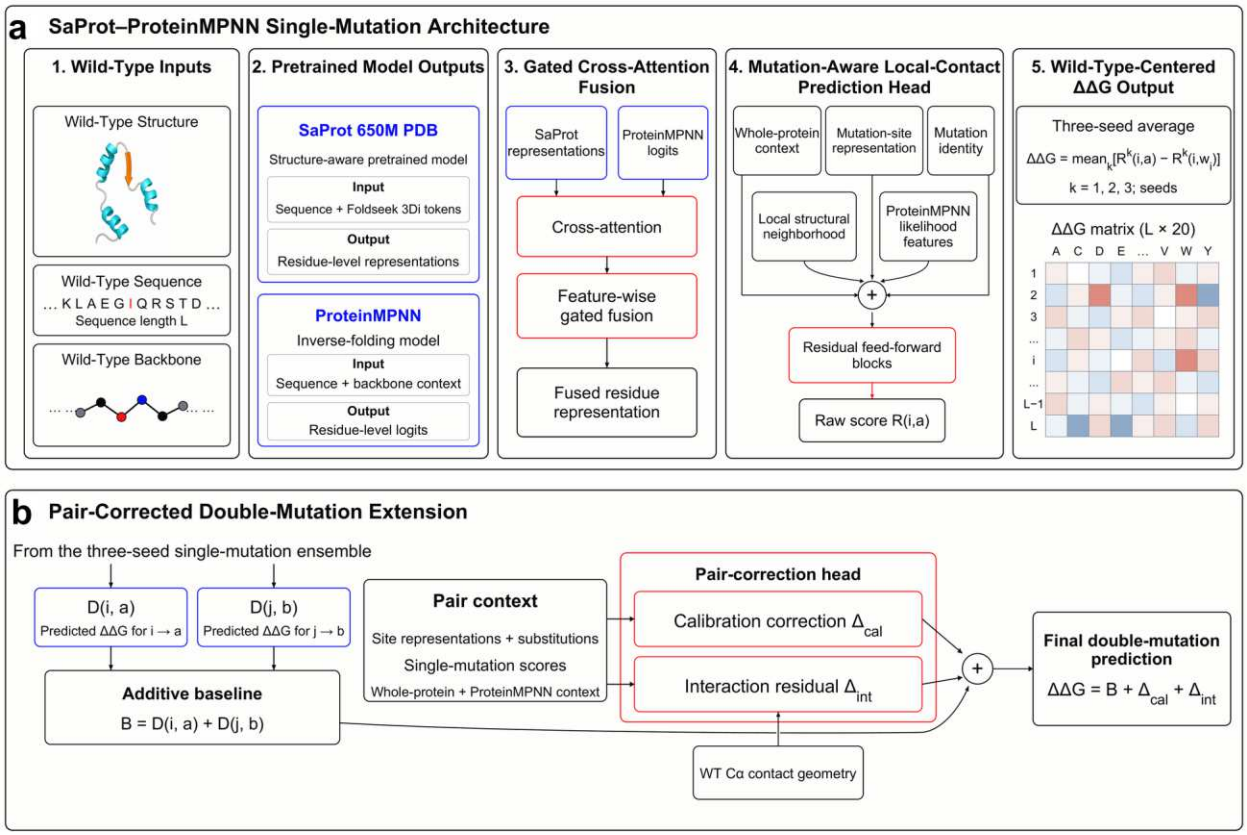



**Figure 1.** SAAFEC-STAIR single- and double-mutation prediction architectures. **a**, The selected SaProt–ProteinMPNN single-mutation model combines frozen SaProt residue representations and ProteinMPNN logits through trainable gated cross-attention. The mutation-aware head integrates the fused mutation-site representation with whole-protein context, local structural neighborhood, mutation identity, and ProteinMPNN likelihood features to produce wild-type-centered ΔΔG predictions. Prediction matrices from three independently initialized models are averaged to form the final L × 20 matrix. **b**, The pair-corrected double-mutation extension reuses the frozen three-seed single-mutation ensemble. Two single-site predictions define an additive baseline; the pair head adds a calibration correction and a contact-modulated interaction residual informed by symmetric pair context and wild-type Cα geometry. Only the double-mutation head is trained.

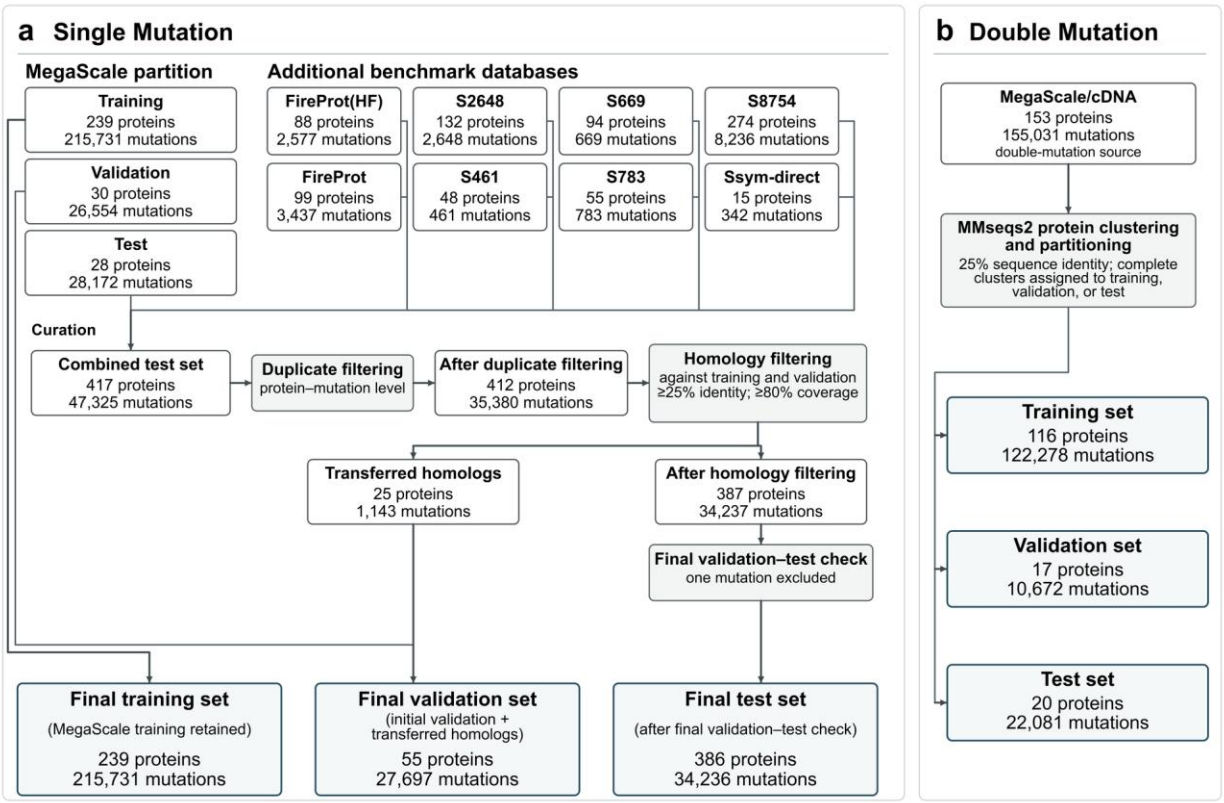



**Figure 2.** Data curation and partitioning of the SAAFEC-STAIR single- and double-mutation datasets. **a**, For the single-mutation workflow, the MegaScale test split was combined with eight additional benchmark databases, followed by protein-mutation-level duplicate filtering and sequence-homology filtering against the MegaScale training and validation sets. Homologous benchmark proteins were transferred to validation, and a final validation-test check produced the held-out test set. **b**, For the double-mutation workflow, we used the ThermoMPNN protein-level partition of the MegaScale/cDNA measurements, in which proteins were clustered at 25% sequence identity and complete clusters were assigned to the training, validation, and test splits.

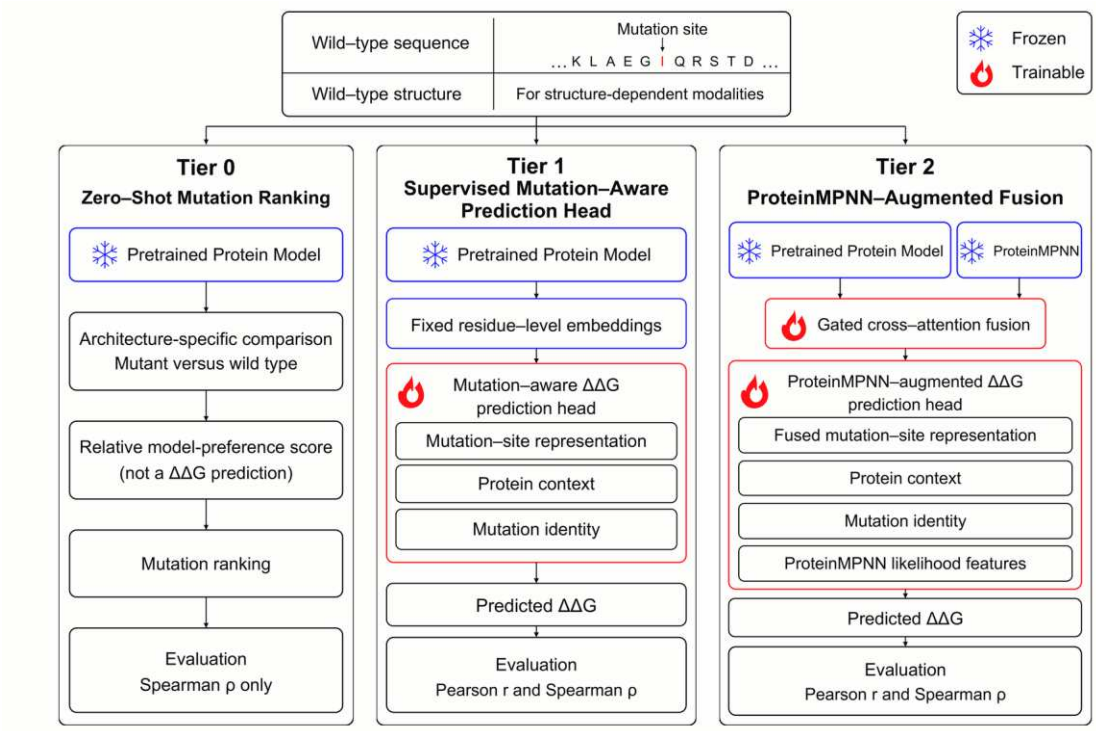



**Figure 3.** Tiered single-mutation modeling framework. Tier 0 uses architecture-specific relative-preference scoring from a frozen pretrained protein model, without ΔΔG supervision, to rank mutations and is evaluated using Spearman ρ only. Tier 1 trains a mutation-aware ΔΔG prediction head on experimental ΔΔG measurements, whereas Tier 2 retains this supervised framework and adds frozen ProteinMPNN information through trainable gated cross-attention fusion. Wild-type structure is supplied only for structure-dependent modalities; snowflake and flame symbols denote frozen and trainable components, respectively.

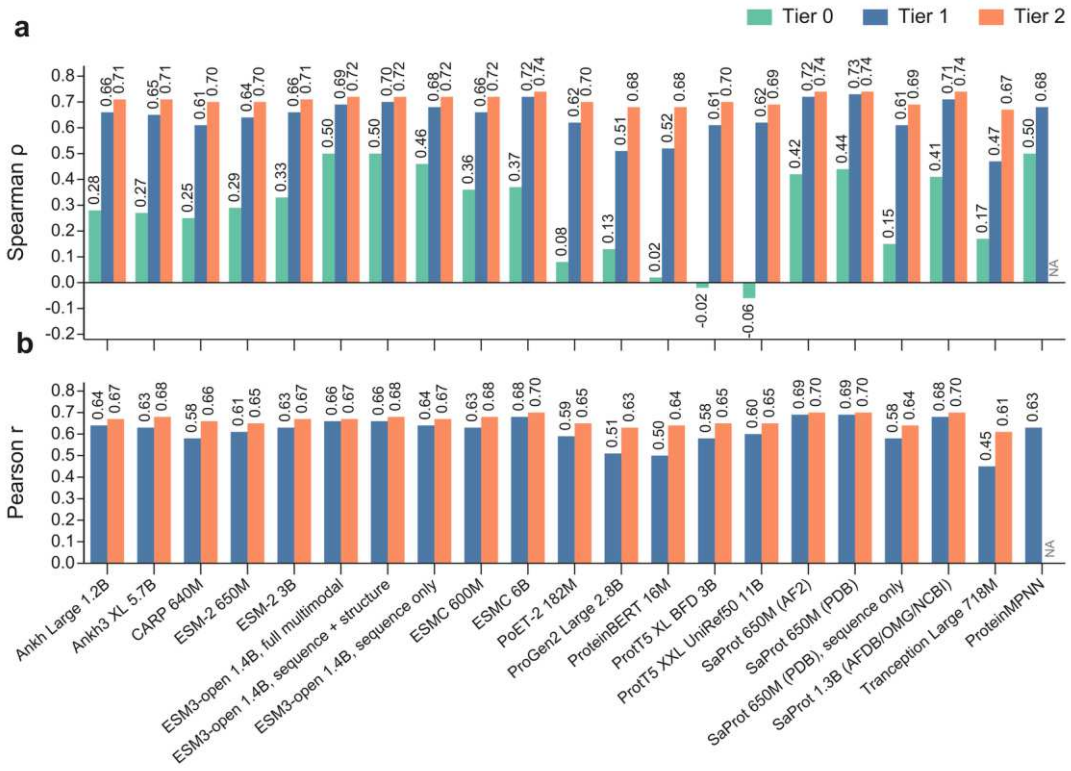



**Figure 4.** Model-wise performance across the tiered single-mutation framework on the final single-mutation test set. **a**, Spearman rank correlation coefficient (ρ) for Tiers 0–2. **b**, Pearson correlation coefficient (r) for Tiers 1 and 2. Tier 0 Pearson correlation is not shown because Tier 0 produces architecture-specific mutation-ranking scores rather than calibrated ΔΔG predictions. Tier 2 adds ProteinMPNN information to each applicable pretrained-model configuration. Values are labeled above the bars; NA denotes a model–tier combination that was unavailable or not applicable.

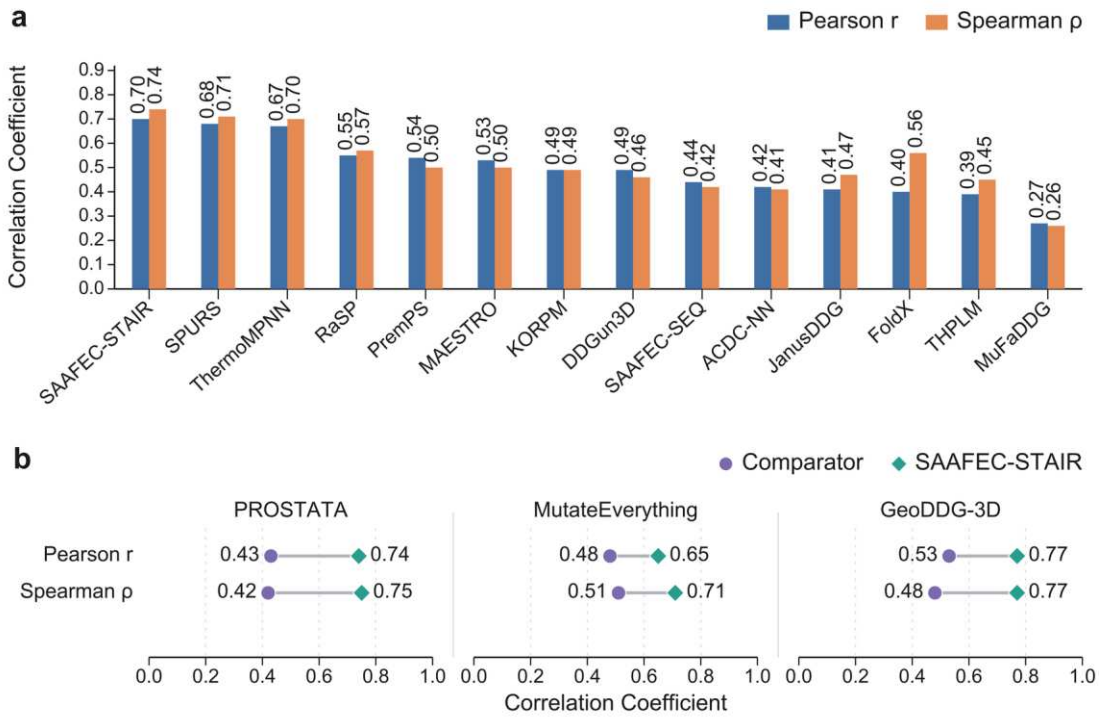



**Figure 5.** Single-mutation predictor comparisons. **a**, Pearson and Spearman correlations on the final test set. **b**, Matched SAAFEC-STAIR and comparator correlations on three method-specific homology-filtered subsets. Panel b comparisons are valid only within each subset.

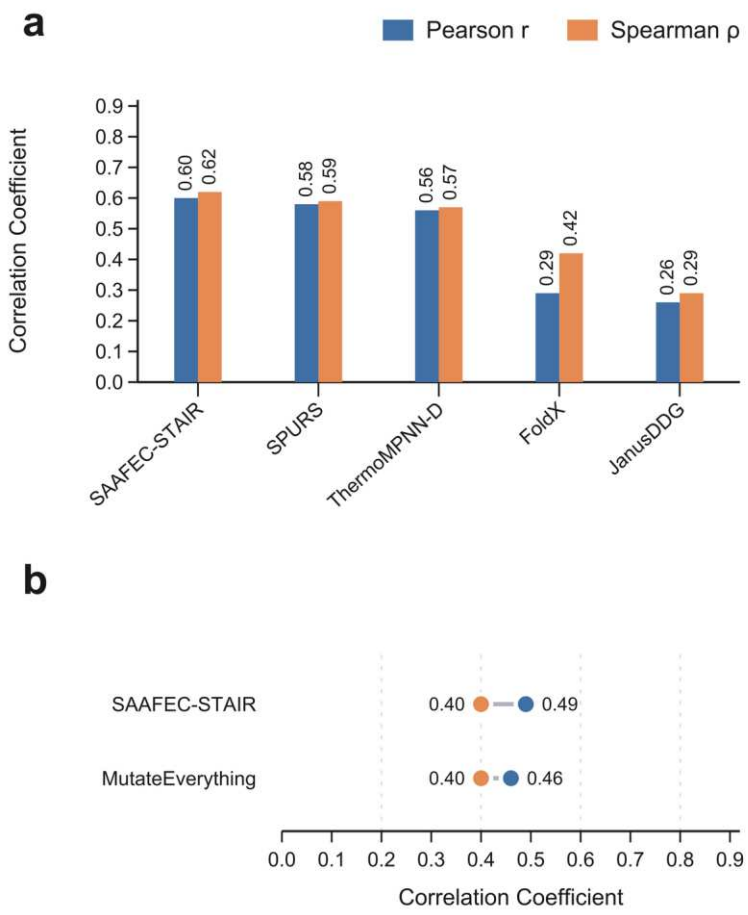



**Figure 6.** Double-mutation predictor comparisons. **a**, Pearson and Spearman correlations on the complete test set. **b**, Matched correlations for SAAFEC-STAIR and MutateEverything on the seven-protein overlap-control subset.

## Tables

**Table 1.** Single-mutation benchmark performance on the final test set (386 proteins; 34,236 mutations).

| Method | Pearson r | Spearman ρ |
|---|---|---|
| SAAFEC-STAIR | 0.70 | 0.74 |
| SPURS | 0.68 | 0.71 |
| ThermoMPNN | 0.67 | 0.70 |
| RaSP | 0.55 | 0.57 |
| PremPS | 0.54 | 0.50 |
| MAESTRO | 0.53 | 0.50 |
| KORPM | 0.49 | 0.49 |
| DDGun3D | 0.49 | 0.46 |
| SAAFEC-SEQ | 0.44 | 0.42 |
| ACDC-NN | 0.42 | 0.41 |
| JanusDDG | 0.41 | 0.47 |
| FoldX | 0.40 | 0.56 |
| THPLM | 0.39 | 0.45 |
| MuFaDDG | 0.27 | 0.26 |

**Table 2.** Double-mutation benchmark performance on the complete test set (20 proteins; 22,081 double mutations).

| Method | Pearson r | Spearman ρ |
|---|---|---|
| SAAFEC-STAIR (pair-corrected) | 0.60 | 0.62 |
| SAAFEC-STAIR (additive single-mutation baseline) | 0.56 | 0.60 |
| SPURS | 0.58 | 0.59 |
| ThermoMPNN-D | 0.56 | 0.57 |
| FoldX | 0.29 | 0.42 |
| JanusDDG | 0.26 | 0.29 |

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SAAFECSTAIRSupplementaryInformation.pdf