Supplementary Information

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

Contents

Supplementary Methods: Data sources and dataset construction 3
S1.1. Single-mutation data sources and dataset construction 3
S1.2. Double-mutation dataset construction and partitioning 5
Supplementary Note: Tiered evaluation framework and results 6
S2.1. Tiered single-mutation modeling framework 6
S2.2. Tiered single-mutation performance and model selection 6
S2.3. PoET-2 Tier 0 scoring-condition analysis 7
Supplementary Note: Benchmark comparison results 8
S3.1. Single-mutation homology-filtered predictor comparisons 8
S3.2. Double-mutation overlap control 8
Supplementary Methods: Model and evaluation details 8
S4.1. Pretrained protein models and common input preparation 8
S4.2. Selected single-mutation architecture details 11
S4.2.1. Structural neighborhood and ProteinMPNN entropy features 11
S4.2.2. Interactive visualization of single-mutation predictions 12
S4.3. Pair-corrected double-mutation architecture 12
S4.4. Training and model-selection settings 13
S4.5. Evaluation metrics and comparison controls 13
S4.5.1. Predictor descriptions and availability 13
S4.5.2. Sign-convention alignment 15
S4.5.3. Single-mutation homology controls 16
Supplementary references 16

Supplementary Methods: Data sources and dataset construction

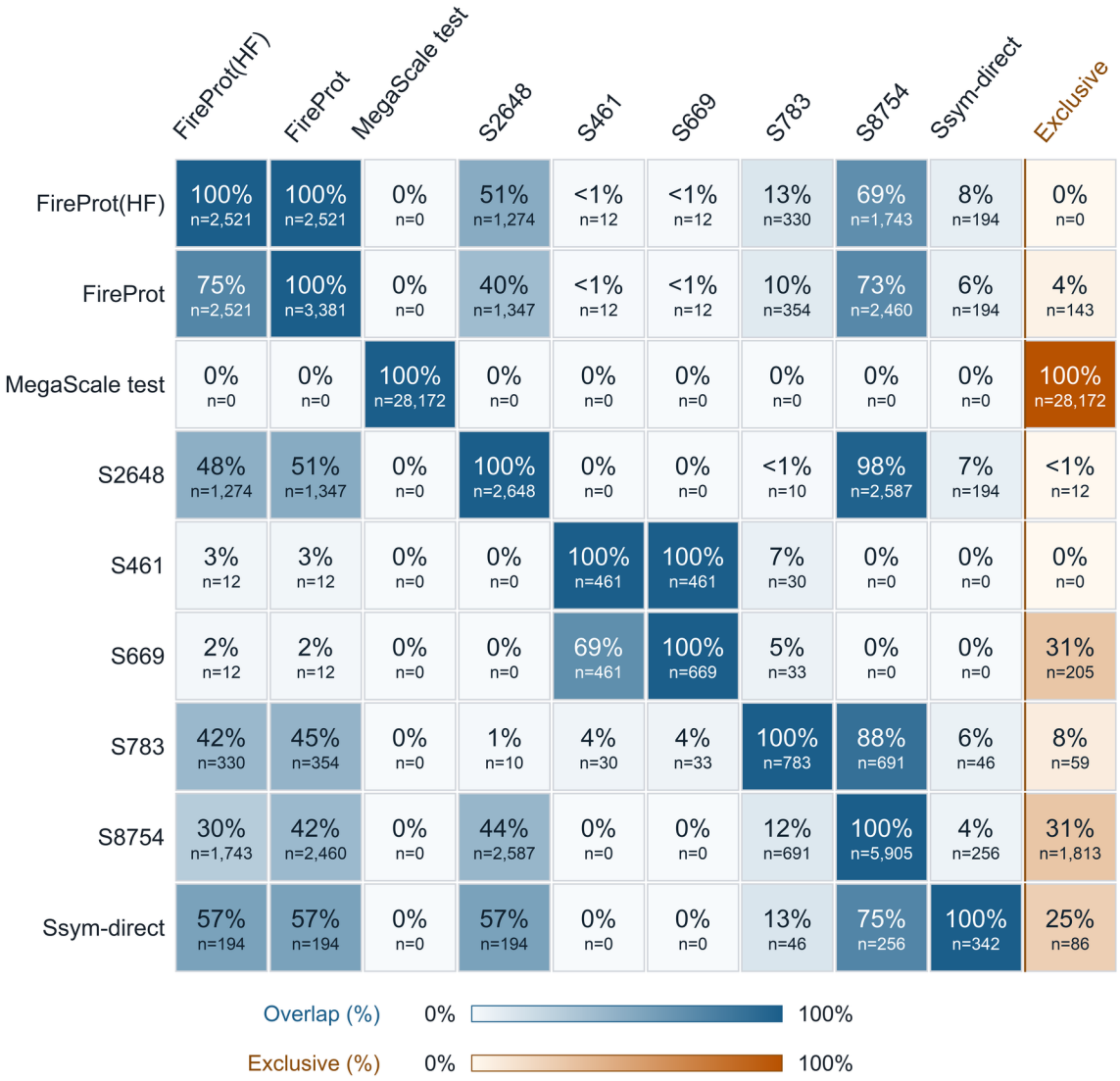
S1.1. Single-mutation data sources and dataset construction

Supplementary Table S1. Single-mutation source databases and dataset sizes before duplicate filtering. Counts refer to measurement records before duplicate filtering; Supplementary Fig. S1 counts unique protein–mutation pairs.

Database	Proteins, n	Measurements, n
MegaScale test	28	28,172
FireProt(HF)	88	2,577
FireProt	99	3,437
S2648	132	2,648
S461	48	461
S669	94	669
S783	55	783
S8754	274	8,236
Ssym-direct	15	342

Supplementary Table S2. Data resources for the single-mutation datasets.

Dataset	Data resource
MegaScale [1]	Zenodo: https://doi.org/10.5281/zenodo.7992926
MegaScale partition [2]	ThermoMPNN dataset splits: https://github.com/Kuhlman-Lab/ThermoMPNN/tree/main/dataset_splits
FireProt and FireProt(HF) [2, 3]	FireProtDB original source and database description: https://doi.org/10.1093/nar/gkaa981
S2648 [4, 5]	INPS dataset page: https://inpsmd.biocomp.unibo.it/datasets/
S669 [6]	Zenodo: https://doi.org/10.5281/zenodo.7568094
S461 [7]	KORPM repository: https://github.com/chaconlab/korpm
S783 and S8754 [8]	GeoStab repository: https://github.com/Gonglab-THU/GeoStab
Ssym-direct [9]	Supplementary data accompanying Pucci et al.: https://academic.oup.com/bioinformatics/article/34/21/3659/4987874



Supplementary Figure S1. Protein-mutation overlap across single-mutation databases before duplicate filtering. Rows represent the queried database and columns represent the comparison database. Each cell reports the percentage of unique protein-mutation pairs in the row database that are also present in the column database, with the number of shared pairs shown below. The rightmost column reports pairs unique to each database.

The filtering stages summarized in Supplementary Table S3 began with a training–validation homology check. Validation protein HHH_rd4_0395_A matched training protein HHH_rd1_0516_A at 45.9% sequence identity and was removed before the final benchmark set was assembled.

Duplicate resolution of the combined benchmark sources followed the consistency criteria described in the main Methods. We averaged 4,894 duplicate groups directly and retained another 372 groups after removing one outlier; 432 groups that remained inconsistent were excluded. Each retained group was represented by the mean of its included measurements. These are counts of duplicate groups; Supplementary Table S3 reports the resulting changes in measurement counts.

Following duplicate resolution, sequence-homology filtering transferred 25 proteins and 1,143 mutations from the benchmark test pool to validation. The remaining test proteins were then checked against this expanded validation set. Test protein 1AZP_A was homologous to validation proteins 1BNZ_A and 1SSO_A. Its only test mutation, V30I, was excluded, producing the final test set reported in Supplementary Table S3.

Supplementary Table S3. Construction and filtering of the single-mutation datasets.

Stage	Proteins, n	Mutations, n	Outcome
MegaScale training set	239	215,731	Used for training
MegaScale validation set after training–validation homology check	30	26,554	One training-homologous validation protein removed
Combined test set	417	47,325	Before duplicate filtering
After duplicate filtering	412	35,380	Duplicate consolidation through averaging reduced the record count by 10,106; 372 outlier measurements were removed separately; 1,467 measurements in 432 unresolved groups were excluded.
After homology filtering	387	34,237	25 proteins and 1,143 mutations transferred to validation
Final validation set	55	27,697	30 MegaScale and 25 transferred test proteins
Final test set	386	34,236	After excluding mutation V30I in 1AZP_A

S1.2. Double-mutation dataset construction and partitioning

The double-mutation files were obtained from the ThermoMPNN-D data archive [10] at https://doi.org/10.5281/zenodo.13345274. The measurements originated from MegaScale [1], and the protein-level partitioning followed ThermoMPNN [2].

Supplementary Table S4. Protein-level partition sizes and measurement count for the double-mutation dataset.

Split	Proteins, n	Double-mutation measurements, n
Training	116	122,278
Validation	17	10,672
Test	20	22,081
Total	153	155,031

Supplementary Note: Tiered evaluation framework and results

S2.1. Tiered single-mutation modeling framework

Supplementary Table S5. Definitions and evaluation roles of the three single-mutation modeling tiers.

Tier	Name	Representation and training	Evaluation role
0	Zero-shot mutation ranking	Frozen, architecture-native amino-acid preference or sequence-likelihood score; no supervised ΔΔG head	Mutation ranking only; evaluated by Spearman ρ only
1	Supervised mutation-aware prediction	Frozen, architecture-native residue-level representations + common supervised mutation-specific readout	Supervised ΔΔG prediction from one pretrained model; evaluated by Pearson r and Spearman ρ
2	ProteinMPNN-augmented fusion	Tier 1 representations + frozen ProteinMPNN logits + common gated cross-attention readout	Effect of adding sequence- and backbone-conditioned ProteinMPNN information; evaluated by Pearson r and Spearman ρ

S2.2. Tiered single-mutation performance and model selection

Supplementary Table S6. Model-wise performance underlying main Fig. 4 on the final single-mutation test set across pretrained models and input configurations in Tiers 0–2. Only Spearman correlations are reported for Tier 0 because Tier 0 produces mutation-ranking scores rather than calibrated ΔΔG predictions. Model-specific inputs and Tier 0 scoring procedures are defined in Supplementary Section S4.1 and Supplementary Table S10; Tier 2 additionally incorporates ProteinMPNN.

Model or input configuration	Tier 0 Spearman ρ	Tier 1 Spearman ρ	Tier 2 Spearman ρ	Tier 1 Pearson r	Tier 2 Pearson r
ProteinMPNN	0.50	0.68	—	0.63	—
Ankh Large 1.2B	0.28	0.66	0.71	0.64	0.67
Ankh3 XL 5.7B	0.27	0.65	0.71	0.63	0.68
CARP 640M	0.25	0.61	0.70	0.58	0.66
ESM-2 650M	0.29	0.64	0.70	0.61	0.65
ESM-2 3B	0.33	0.66	0.71	0.63	0.67
ESMC 600M	0.36	0.66	0.72	0.63	0.68
ESMC 6B	0.37	0.72	0.74	0.68	0.70
PoET-2 182M	0.08	0.62	0.70	0.59	0.65

Model or input configuration	Tier 0 Spearman ρ	Tier 1 Spearman ρ	Tier 2 Spearman ρ	Tier 1 Pearson r	Tier 2 Pearson r
ProGen2 Large 2.8B	0.13	0.51	0.68	0.51	0.63
ProteinBERT 16M	0.02	0.52	0.68	0.50	0.64
ProtT5 XL BFD 3B	−0.02	0.61	0.70	0.58	0.65
ProtT5 XXL UniRef50 11B	−0.06	0.62	0.69	0.60	0.65
Tranception Large 718M	0.17	0.47	0.67	0.45	0.61
ESM3-open 1.4B, sequence only	0.46	0.68	0.72	0.64	0.67
ESM3-open 1.4B, sequence + structure	0.50	0.70	0.72	0.66	0.68
ESM3-open 1.4B, full multimodal	0.50	0.69	0.72	0.66	0.67
SaProt 650M (PDB), sequence only	0.15	0.61	0.69	0.58	0.64
SaProt 1.3B (AFDB/OMG/NCBI)	0.41	0.71	0.74	0.68	0.70
SaProt 650M (AF2)	0.42	0.72	0.74	0.69	0.70
SaProt 650M (PDB)	0.44	0.73	0.74	0.69	0.70

S2.3. PoET-2 Tier 0 scoring-condition analysis

PoET-2 was evaluated under three conditions to distinguish the contribution of the scoring formulation from that of homologous sequences (Supplementary Table S7). Query-sequence-only causal scoring was used for the primary cross-model result in main Fig. 4 and Supplementary Table S6. The homolog-context condition was reported separately because it supplied additional evolutionary information.

Supplementary Table S7. PoET-2 Tier 0 scoring conditions. All values are global Spearman rank correlations with experimental ΔΔG on the final single-mutation test set.

Condition	Scoring procedure and information supplied	Spearman ρ	Reporting role
Sequence causal (query only)	Complete wild-type–mutant causal sequence likelihood; query sequence only	0.08	Primary Tier 0 value; main Fig. 4 and Supplementary Table S6
Masked-residue scoring (query only)	Target residue masked; bidirectional sequence context; query sequence only	0.18	Supplementary alternative scoring condition
MSA causal (query + homologs)	Causal sequence likelihood with retrieved homolog sequences supplied as context	0.37	Supplementary homolog-context analysis

Supplementary Note: Benchmark comparison results

S3.1. Single-mutation homology-filtered predictor comparisons

Supplementary Table S8. Matched performance on method-specific homology-filtered subsets of the final single-mutation test set. Each row represents a different filtered subset.

Comparator	Mutations retained, n	Comparator Pearson r	SAAFEC-STAIR Pearson r	Comparator Spearman ρ	SAAFEC-STAIR Spearman ρ
PROSTATA	8,999	0.43	0.74	0.42	0.75
MutateEverything	15,435	0.48	0.65	0.51	0.71
GeoDDG-3D	24,128	0.53	0.77	0.48	0.77

S3.2. Double-mutation overlap control

For the double-mutation comparison with MutateEverything, 13 of the 20 test-protein identifiers appeared in the published cDNA training split, and 15,374 of the 22,081 double-mutation records were exact overlaps. The control removed all records from those 13 proteins, including records that were not exact overlaps, leaving seven proteins and 4,599 mutations.

Supplementary Table S9. Matched performance on the seven-protein double-mutation overlap-control subset (4,599 double mutations).

Method	Pearson r	Spearman ρ
SAAFEC-STAIR (pair-corrected)	0.49	0.40
MutateEverything	0.46	0.40

Supplementary Methods: Model and evaluation details

S4.1. Pretrained protein models and common input preparation

Supplementary Table S10 summarizes the scale, inference inputs, architecture, and Tier 0 scoring procedure for each pretrained-model configuration. Information supplied during inference is distinguished from information used only during pretraining; checkpoints from the same model family are grouped only when these properties are shared. At inference, sequence-only models received only the amino-acid sequence; the SaProt sequence-only ablation omitted 3Di tokens. The structure-aware SaProt configurations received residue-aligned amino-acid and 3Di tokens [11], whereas the ESM3 configurations received the sequence, structure, secondary-structure (SS8), and solvent-accessibility (SASA) tracks specified for that condition [12]. Sequence-only Tier 1 branches did not receive separate PDB-derived local-contact features.

ProteinMPNN logits were generated with the ThermoMPNN implementation [2] and the official ProteinMPNN checkpoint [13], using the wild-type backbone and full wild-type amino-acid sequence, and were supplied to every applicable Tier 2 branch. The same frozen logits were used with the common supervised mutation-aware head for the standalone ProteinMPNN Tier 1 baseline. This baseline was distinct from the published ThermoMPNN [2] predictor evaluated in the single-mutation benchmark.

Tier 0 scoring was adapted to each model architecture. For masked encoders such as ESM-2, ESMC, and CARP, the target residue was hidden and wild-type-minus-mutant log-probability differences were calculated using sequence context on both sides [14, 15, 16]. The structure-aware SaProt configurations used the same bidirectional approach but masked only the amino-acid component while retaining the wild-type 3Di token [11]. For the encoder–decoder models ProtT5, Ankh Large, and Ankh3 XL, the target residue was replaced by the model-specific sentinel token, a placeholder marking the residue to be reconstructed. Decoder probabilities for the wild-type and mutant residues were then compared in 32-bit floating-point precision [17, 18, 19, 20].

The autoregressive models ProGen2 and Tranception score sequences one residue at a time and were evaluated by comparing the complete wild-type and mutant sequence likelihoods in both forward and reverse orientations and averaging the two differences [21, 22]. Among the evaluated models, only Tranception and PoET-2 offered options to supply homologous sequences during inference. The primary cross-model comparison used the query sequence alone for both models: Tranception's optional retrieval component was disabled [22], and PoET-2 used its query-sequence-only causal score [23]. The additional PoET-2 scoring conditions and results are reported in Supplementary Table S7.

For ESM3, the target sequence token was masked while the other wild-type tracks supplied for that condition were held fixed. Wild-type and mutant probabilities from the sequence track were used to calculate wild-type-minus-mutant log-probability differences [12].

ProteinBERT differs from the other evaluated models because its pretraining jointly used amino-acid sequence modeling and Gene Ontology (GO) annotation prediction [24]. GO annotations are standardized whole-protein functional labels covering molecular function, biological process, and cellular component. The GO input vector was set to zero, following the original ProteinBERT benchmark protocol [24], so no protein-specific functional annotations were supplied. Tier 0 used the model's native residue-denoising score averaged across the 26 token-replacement evaluations at the target site, whereas Tiers 1 and 2 used the frozen sequence-derived representations.

Supplementary Table S10. Pretrained models, approximate parameter counts, inputs, architectures, and Tier 0 scoring procedures.

Model or configuration	Approximate parameters	Inputs	Architecture	Tier 0 scoring
ESM-2	650M; 3B	Sequence only	Masked encoder	Target-residue masking; wild-type–mutant log-probability difference
ESMC	600M; 6B	Sequence only	ESM Cambrian masked encoder	Target-residue masking; wild-type–mutant log-probability difference
ProtT5 (XL/XXL)	3B; 11B	Sequence only	T5 encoder–decoder	Single-residue sentinel-token reconstruction
Ankh Large	1.2B	Sequence only	Protein T5-like encoder–decoder	Single-residue sentinel-token reconstruction

Model or configuration	Approximate parameters	Inputs	Architecture	Tier 0 scoring
Ankh3 XL	5.7B	Sequence only	Encoder–decoder trained by sequence denoising and completion	Single-residue sentinel-token reconstruction (denoising mode)
CARP	640M	Sequence only	Convolutional/ByteNet masked model	Target-residue masking; wild-type–mutant log-probability difference
ProGen2 Large	2.8B	Sequence only	Autoregressive sequence model	Complete wild-type–mutant sequence likelihood averaged over forward and reversed sequences
Tranception Large	718M	Sequence only	Autoregressive sequence model	Complete wild-type–mutant sequence likelihood averaged over forward and reversed orientations; no homolog retrieval
PoET-2	182M	Sequence only	Hierarchical encoder with causal and masked decoders	Primary: query-only causal wild-type–mutant sequence likelihood
ProteinBERT	16M	Sequence only; GO input set to zero	Compact local-global Transformer	Target-site residue-denoising score; GO input set to zero
SaProt	650M; 1.3B	Amino acid + Foldseek/3Di structure tokens; sequence-only condition also tested	ESM-style masked model over paired amino-acid and structural tokens	Amino-acid component masked; wild-type 3Di token retained in structure-aware conditions and omitted in the sequence-only ablation
ESM3-open	1.4B	Sequence only; sequence + structure; full multimodal (sequence + structure + SS8 + SASA)	Multitrack generative model	Sequence token masked; conditioned on the tracks supplied in each configuration
ProteinMPNN	≈1.7M	Wild-type backbone geometry + wild-type amino-acid sequence	Message-passing inverse-folding model	Sequence- and backbone-conditioned wild-type–mutant log-probability difference

Supplementary Table S11. Code repositories and model-weight sources for the pretrained protein models evaluated in this study.

Model (parameters)	Code repository	Model-weight source
ESM-2 (650M)	https://github.com/facebookresearch/esm	https://huggingface.co/facebook/esm2_t33_650M_UR50D
ESM-2 (3B)	https://github.com/facebookresearch/esm	https://huggingface.co/facebook/esm2_t36_3B_UR50D
ESMC (600M)	https://github.com/Biohub/esm	https://huggingface.co/biohub/esmc-600m-2024-12

Model (parameters)	Code repository	Model-weight source
ESMC (6B)	https://github.com/Biohub/esm	https://huggingface.co/biohub/esmc-6b-2024-12
ProtT5 XL (3B)	https://github.com/agemagician/ProtTrans	https://huggingface.co/Rostlab/prot_t5_xl_bfd
ProtT5 XXL (11B)	https://github.com/agemagician/ProtTrans	https://huggingface.co/Rostlab/prot_t5_xxl_uniref50
Ankh Large (1.2B)	https://github.com/agemagician/Ankh	https://huggingface.co/ElnaggarLab/ankh-large
Ankh3 XL (5.7B)	https://github.com/agemagician/Ankh	https://huggingface.co/ElnaggarLab/ankh3-xl
CARP (640M)	https://github.com/microsoft/protein-sequence-models	https://zenodo.org/records/6564798/files/carp_640M.pt
ProGen2 Large (2.8B)	https://github.com/salesforce/progen	https://storage.googleapis.com/sfr-progen-research/checkpoints/progen2-large.tar.gz
Tranception Large (718M)	https://github.com/OATML-Markslab/Tranception	https://huggingface.co/OATML-Markslab/Tranception_Large
PoET-2 (182M)	https://github.com/OpenProteinAI/PoET-2	https://zenodo.org/records/16815645/files/poet-2.ckpt
ProteinBERT (16M)	https://github.com/nadavbra/protein_bert	https://github.com/nadavbra/proteinbert_data_files/blob/master/epoch_92400_sample_23500000.pkl
SaProt (650M; PDB)	https://github.com/westlake-repl/SaProt	https://huggingface.co/westlake-repl/SaProt_650M_PDB
SaProt (650M; AF2)	https://github.com/westlake-repl/SaProt	https://huggingface.co/westlake-repl/SaProt_650M_AF2
SaProt (1.3B)	https://github.com/westlake-repl/SaProt	https://huggingface.co/westlake-repl/SaProt_1.3B_AFDB_OMG_NCBI
ESM3-open (1.4B)	https://github.com/Biohub/esm	https://huggingface.co/biohub/esm3-sm-open-v1
ProteinMPNN (approximately 1.7M)	https://github.com/dauparas/ProteinMPNN	https://github.com/dauparas/ProteinMPNN/blob/main/vanilla_model_weights/v_48_020.pt

S4.2. Selected single-mutation architecture details

S4.2.1. Structural neighborhood and ProteinMPNN entropy features

For the single-mutation head, let d_{ij} denote the wild-type Cα distance between residues *i* and *j*, and let α_{ij} denote the weight assigned to residue *j* when computing the neighborhood around residue *i*. The neighborhood weight is defined as

$$\alpha_{ij} = \begin{cases} \exp\left(-\frac{d_{ij}}{4\,\text{Å}}\right) & \text{if } i \neq j \text{ and } d_{ij} \leq 10\,\text{Å} \\ 0 & \text{otherwise} \end{cases} \tag{S1a}$$

Let W^N denote the learned projection into the 768-dimensional neighborhood-context space, and let $\varepsilon = 10^{-8}$ provide a numerical lower bound for the denominator. Using the fused residue states defined in Eq. (5c) of the main text, the local-neighborhood vector is

$$\boldsymbol{N}_i = \boldsymbol{W}^N \left(\frac{\sum \alpha_{ij} \boldsymbol{F}_j}{\max(\sum \alpha_{ij}, \varepsilon)} \right) \tag{S1b}$$

At most, the 16 largest nonzero weights are retained. Both sums in Eq. (S1b) run over these retained neighbors. If no valid neighbors are available, the pre-projection average is set to zero. This branch, therefore, summarizes the local structural environment around residue *i* using the fixed wild-type coordinates.

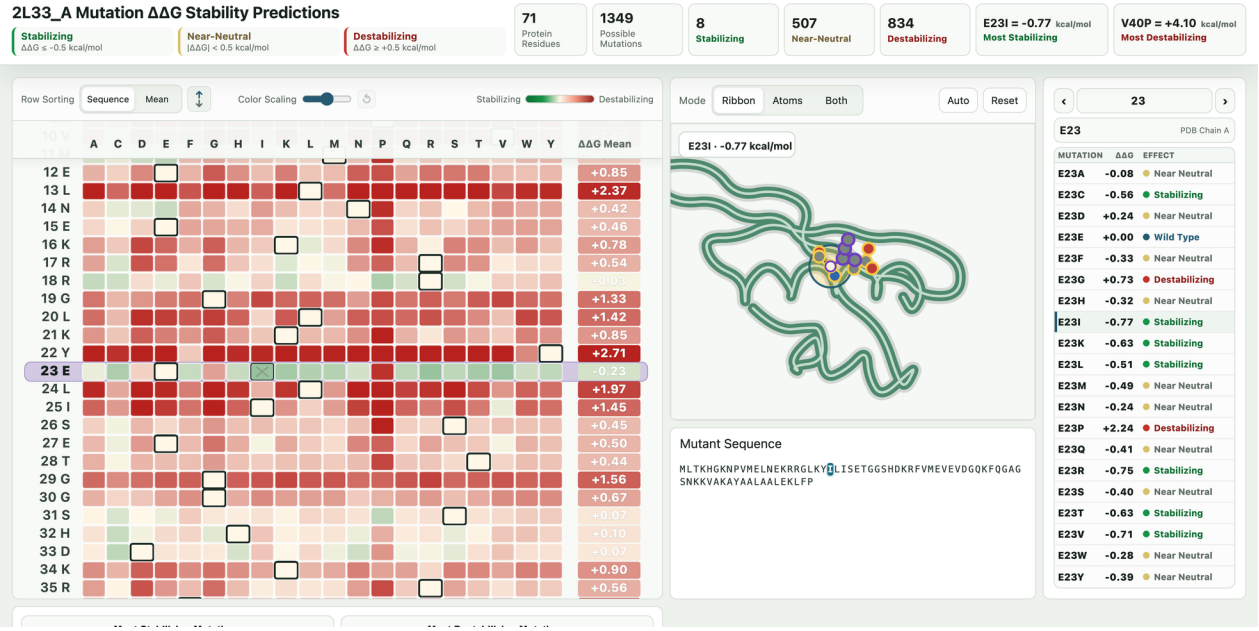
For ProteinMPNN logits Z_i, let $p_i(x) = \mathrm{softmax}(Z_i)_x$ denote the normalized ProteinMPNN-derived probability assigned to amino acid x at residue i. The entropy feature used in the single-mutation head is

$$\mathcal{H}_i = -\sum_{x=1}^{20} p_i(x)\ln\big(p_i(x)\big) \tag{S1c}$$

Lower $\mathcal{H}_i$ indicates that ProteinMPNN concentrates probability on a small set of amino acids, whereas higher $\mathcal{H}_i$ indicates a broader compatibility distribution. The same $\mathcal{H}_i$ value is supplied to every possible substitution at residue i.

S4.2.2. Interactive visualization of single-mutation predictions

The interactive HTML output combines the complete ΔΔG heat map with a rotatable protein-structure view, sequence display, residue-level substitution table, and summaries of stabilizing, near-neutral, and destabilizing predictions. Selections are linked across these views, allowing users to move between residue positions and access ranked lists of the most stabilizing and destabilizing mutations (Supplementary Fig. S2).



Supplementary Figure S2. Illustrative SAAFEC-STAIR interactive HTML interface for exploring the single-mutation prediction matrix.

S4.3. Pair-corrected double-mutation architecture

The double-mutation extension uses predictions from the frozen three-seed single-mutation ensemble as an additive prior. The final prediction $\Delta\Delta G_{ij}^{ab}$ combines the additive baseline B_{ij}^{ab}, calibration term $\Delta_{cal,ij}^{ab}$, and interaction residual $\Delta_{int,ij}^{ab}$, as defined in Eq. (11) of the main text. Only the double-mutation head is trained; the upstream single-mutation ensemble and its pretrained components remain frozen. The interaction residual represents a statistical departure from additivity and should not be interpreted as a direct measurement of epistasis.

S4.4. Training and model-selection settings

Supplementary Table S12. Optimization, validation, early-stopping, and ensemble settings for the single- and double-mutation models.

Setting	Single-mutation models (Tiers 1 and 2)	Double-mutation pair-correction model
Trainable component	Mutation-aware prediction head (Tier 1) or gated cross-attention fusion and prediction head (Tier 2); pretrained models frozen	Calibration and contact-gated pair head; single-mutation ensemble frozen
Optimizer	AdamW	AdamW
Learning rate	1×10^{-3}	3×10^{-4}
Weight decay	0.01	0.01
Batching strategy	1 protein per batch; gradients accumulated over 4 proteins	1 protein per batch; gradients accumulated over 4 proteins
Schedule	5% linear warmup, cosine decay	5% linear warmup, cosine decay
Maximum epochs	100	100
Checkpoint criterion	Best validation performance (Pearson correlation)	Highest global validation Pearson correlation
Early stopping	Patience 10; minimum improvement 0.005	Patience 12; minimum improvement 0.003
Gradient clipping	1.0	1.0
Ensemble	Seeds 1337, 2027, and 3407; predictions averaged after model selection	Fixed three-seed single-mutation ensemble; trained double-mutation head
Test use	Final evaluation after model and ensemble selection	Final evaluation after model selection

S4.5. Evaluation metrics and comparison controls

S4.5.1. Predictor descriptions and availability

The evaluated predictors differed in methodological approach, input requirements, training or parameter sources, and software availability (Supplementary Table S13). The Availability column lists an official GitHub repository or standalone software page. The Input modality column classifies each evaluated configuration as sequence-based or structure-dependent; model-derived representations and descriptors are summarized under the Approach column.

Supplementary Table S13. Existing predictors: methods, inputs, training or parameter sources, references, and software availability.

Method	Approach	Input modality	Training or parameter source	Availability
PROSTATA [25]	Sequence-only ΔΔG predictor using ESM-2 embeddings and an ensemble of five regression heads.	Sequence	S5251, a curated aggregate assembled from established ProTherm/VariBench-derived stability datasets.	GitHub: github.com/AIRI-Institute/PROSTATA

Method	Approach	Input modality	Training or parameter source	Availability
SPURS [26]	Adapter-based fusion of ESM-2 sequence representations with ProteinMPNN structure-derived features.	Structure	Homology-filtered MegaScale training split; MegaScale double-mutation data for the higher-order extension.	GitHub: github.com/luo-group/SPURS
MutateEverything (ESM-2) [27]	ESM-2-based framework using lightweight decoders to predict single- and higher-order mutation effects.	Sequence	MegaScale cDNA training split containing single- and double-mutation measurements.	GitHub: github.com/jozhang97/MutateEverything
ThermoMPNN [2]	ProteinMPNN residue representations followed by a supervised ΔΔG prediction head for single mutations.	Structure	MegaScale single-mutation data.	GitHub: github.com/Kuhlman-Lab/ThermoMPNN
ThermoMPNN-D [10]	Siamese extension of ThermoMPNN for double-mutation prediction using ProteinMPNN residue and pairwise structural features.	Structure	MegaScale-D double-mutation data.	GitHub: https://github.com/Kuhlman-Lab/ThermoMPNN-D
GeoDDG-3D [8]	GeoStab-suite ΔΔG model using a DMS-pretrained geometric encoder with sequence representations and three-dimensional structural information.	Structure	DMS pretraining followed by S8754 ΔΔG supervision.	GitHub: github.com/Gonglab-THU/GeoStab
RaSP [28]	Structure-based neural surrogate trained to approximate Rosetta cartesian_ddg from local three-dimensional residue environments.	Structure	Homology-reduced PDB structures for self-supervised pretraining; Rosetta Cartesian ΔΔG labels for downstream fitting.	GitHub: github.com/KULL-Centre/_2022_ML-ddG-Blaabjerg
PremPS [29]	Structure-based random forest using evolutionary, sequence, and local structural descriptors.	Structure	S5296, a balanced ProTherm-derived dataset containing forward and reverse mutations.	GitHub: github.com/minghuilab/PremPS
MAESTRO [30]	Structure-based multi-agent predictor combining statistical scoring functions, structural descriptors, and machine learning.	Structure	ProTherm-derived ΔΔG data for the learning agents; a nonredundant PDB set for the statistical scoring functions.	Standalone software: pbwww.services.came.sbg.ac.at/?page_id=477
KORPM [7]	Six-dimensional orientation-dependent residue-pair statistical potential derived from structural statistics.	Structure	PDB-derived orientational statistics, with parameters fitted to curated ProTherm-derived ΔΔG data.	GitHub: github.com/chaconlab/korpm

Method	Approach	Input modality	Training or parameter source	Availability
DDGun3D [31]	Untrained hybrid score using sequence/evolutionary information with solvent accessibility and structural-neighbor statistical potentials.	Structure	No supervised ΔΔG training; sequence profiles and fixed statistical potentials.	GitHub: github.com/biofold/ddgun
SAAFEC-SEQ [32]	Sequence-only gradient-boosting predictor using PSSM/PsePSSM, conservation, physicochemical, and local sequence descriptors.	Sequence	S2648 training set.	Standalone/web: compbio.clemson.edu/saafec-seq/
ACDC-NN [33]	Structure-aware antisymmetry-constrained neural predictor using evolutionary profiles and local three-dimensional descriptors.	Structure	Pretraining on simulated DDGun3D scores followed by fitting to S2648 experimental ΔΔG values.	GitHub: github.com/compbiomed-unito/acdc-nn
JanusDDG [34]	Sequence-based, physics-informed PLM predictor with bidirectional attention and constraints for antisymmetry and transitivity.	Sequence	S2450 (filtered from S2648), augmented with reverse mutations.	GitHub: github.com/compbiomed-unito/JanusDDG
FoldX [35]	Empirical structure-based energy function using packing, solvation, hydrogen bonding, electrostatics, entropy, and local repair terms.	Structure	Empirical energy terms calibrated against experimental thermodynamic data.	Standalone software: foldxsuite.crg.es/
THPLM [36]	Sequence-only predictor using ESM-2 embeddings with convolutional and regression layers.	Sequence	S2648 with reverse-mutation augmentation.	GitHub: github.com/FPPGroup/THPLM
MuFaDDG [37]	ESM-2 global sequence features fused with NetSurfP-predicted local structural and physicochemical features.	Sequence	S2648 with homology-aware folds.	GitHub: github.com/PengjiaMa23/MuFaDDG

S4.5.2. Sign-convention alignment

Before evaluation, experimental measurements and predictor outputs were aligned to the ΔΔG sign convention defined by Equation (1) of the main text, in which positive values indicate destabilization. Experimental values were negated for MegaScale, S2648, S461, S669, S783, S8754, and Ssym-direct, whereas FireProt values required no sign change. Predictor outputs were negated for KORPM, DDGun3D, SAAFEC-SEQ, ACDC-NN, JanusDDG, THPLM, MuFaDDG, PROSTATA, and GeoDDG-3D; no sign change was applied to SAAFEC-STAIR, SPURS, ThermoMPNN, ThermoMPNN-D, RaSP, PremPS, MAESTRO, FoldX, or

MutateEverything. Tier 0 score-direction alignment was treated as ranking alignment rather than a thermodynamic conversion.

S4.5.3. Single-mutation homology controls

For PROSTATA, the ESM-2 configuration of MutateEverything, and GeoDDG-3D, method-specific subsets of the final single-mutation test set retained proteins with <25% sequence identity and no detected homology to the corresponding predictor's training proteins. Within each subset, SAAFEC-STAIR and the comparator were evaluated on the same set of retained mutations (Supplementary Table S8).

Supplementary Table S14. Proteins and mutations removed and retained by the method-specific homology filters used for the single-mutation comparisons.

Predictor	Proteins removed, n	Mutations removed, n	Proteins retained, n	Mutations retained, n
PROSTATA	344	25,237	42	8,999
MutateEverything	32	18,801	354	15,435
GeoDDG-3D	287	10,108	99	24,128

Supplementary references

1. Tsuboyama, K. et al. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620,** 434-444 (2023).
2. 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).
3. Stourac, J. et al. FireProtDB: database of manually curated protein stability data. *Nucleic Acids Res.* **49,** D319-D324 (2021).
4. 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).
5. 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).
6. 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).
7. 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).
8. 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).
9. 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).
10. Dieckhaus, H. & Kuhlman, B. Protein stability models fail to capture epistatic interactions of double point mutations. *Protein Sci.* **34,** e70003 (2025).
11. Su, J. et al. SaProt: protein language modeling with structure-aware vocabulary. In: *The Twelfth International Conference on Learning Representations* (2024).
12. Hayes, T. et al. Simulating 500 million years of evolution with a language model. *Science* **387,** 850-858 (2025).

13. Dauparas, J. et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378,** 49-56 (2022).
14. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379,** 1123-1130 (2023).
15. Candido, S. et al. Language modeling materializes a world model of protein biology. *bioRxiv*, https://doi.org/10.64898/2026.06.03.729735 (2026).
16. Yang, K. K., Fusi, N. & Lu, A. X. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Syst.* **15,** 286-294.e2 (2024).
17. 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).
18. 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).
19. Elnaggar, A. et al. Ankh: optimized protein language model unlocks general-purpose modelling. *arXiv*, https://doi.org/10.48550/arXiv.2301.06568 (2023).
20. 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).
21. 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).
22. 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).
23. 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).
24. 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).
25. Umerenkov, D. et al. PROSTATA: a framework for protein stability assessment using transformers. *Bioinformatics* **39,** btad671 (2023).
26. Li, Z. & Luo, Y. Generalizable and scalable protein stability prediction with rewired protein generative models. *Nat. Commun.* **17,** 891 (2026).
27. 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).
28. Blaabjerg, L. M. et al. Rapid protein stability prediction using deep learning representations. *eLife* **12,** e82593 (2023).
29. Chen, Y. et al. PremPS: predicting the impact of missense mutations on protein stability. *PLoS Comput. Biol.* **16,** e1008543 (2020).
30. Laimer, J., Hofer, H., Fritz, M., Wegenkittl, S. & Lackner, P. MAESTRO - multi agent stability prediction upon point mutations. *BMC Bioinformatics* **16,** 116 (2015).
31. 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).
32. 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).
33. 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).
34. Barducci, G. et al. JanusDDG: a physics-informed neural network for sequence-based protein stability via two-fronts attention. *Commun. Biol.* **9,** 494 (2026).

35. 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).
36. 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).
37. Gong, J. et al. MuFaDDG: a sequence-based multiscale feature fusion framework for protein stability changes prediction. *Bioinformatics* **42,** btag196 (2026).

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.