Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

Published on 11 August 2026

Cite this: DOI: 10.1039/d6sc02103d

All publication charges for this article have been paid for by the Royal Society of Chemistry

A unified predictor of protein stability changes across all mutation types *via* implicit structure learning

Hong Tan, [ab] Shenggeng Lin[abc] and Yi Xiong *[abc]

Received 13th March 2026
Accepted 21st July 2026

DOI: 10.1039/d6sc02103d

rsc.li/chemical-science

Prediction of protein stability change caused by amino acid substitutions or indels (insertions/deletions) is crucial for protein engineering. While current models excel at single-point substitutions, they struggle with multi-point mutations and indels due to simplistic additivity assumptions and the inability to model backbone conformational changes. To address these limitations, we introduce UniStab, an end-to-end framework for predicting stability changes across all mutation types. By leveraging the implicit geometric reasoning of a pre-trained folding model, UniStab effectively captures non-additive epistatic interactions and local backbone rearrangements without the prohibitive cost of explicit structure generation. Evaluated on a comprehensive benchmark, UniStab demonstrates state-of-the-art performance, particularly in the challenging scenarios of multi-point mutations and indels. Beyond predictive accuracy, UniStab provides interpretable structural insights and effectively guides the design of stabilized variants, facilitating its potential utility in rational protein engineering.

1 Introduction

Protein stability is fundamental to biological function by maintaining a protein's folded structure, and it is a prerequisite for designing robust industrial enzymes and next-generation therapeutics.[1–6] In energetic terms, stability reflects the free-energy difference (ΔG) between folded and unfolded states.[7,8] The effect of a mutation is commonly quantified as the change in stability, $\Delta\Delta G$, which distinguishes stabilizing from destabilizing variants. Although high-throughput approaches (*i.e.*, deep mutational scanning) have begun to map these landscapes, experimental measurements remain condition-specific and costly, covering only a tiny fraction of the protein universe.[9–12] Consequently, for the vast majority of possible sequence variants, the thermodynamic impact of mutation remains unknown.[11]

Continuous efforts have been made to develop computational methods for reducing experimental burdens in protein engineering.[13] Existing computational approaches can be broadly grouped into two categories: physics-based methods and data-driven methods (*i.e.*, deep learning-based methods developed in recent years). Physics-based methods estimate stability changes using energy-based scoring functions, with representative tools including FoldX[14] and Rosetta.[15] Data-driven methods infer stability changes from data and can be further distinguished by whether or how structural information is incorporated, which include sequence-based and structure-based methods. Sequence-based deep learning models derive predictions solely from primary sequences, often leveraging pre-trained protein language models.[16–18] For example, PROSTATA[16] uses ESM2 (ref. 19) to extract residue-level representations for the wild-type and mutant sequences, and captures mutational effects *via* representation fusion operations such as subtraction or concatenation. Structure-based deep learning approaches can be broadly categorized into four groups. The first encodes the static wild-type microenvironment, formulating stability change prediction as supervised learning conditioned on local structural features.[20,21] For instance, StabilityOracle contrasts learned representations of wild-type and mutant residues within the local context to predict $\Delta\Delta G$.[20] The second incorporates structural signals through predicted geometric intermediates.[22,23] For instance, SPIRED-Stab integrates C_α coordinates and confidence estimates (pLDDT) with ESM2 sequence embeddings to reason about stability change.[22] A third class utilizes inverse folding models, originally developed for sequence design, which assess residue–backbone compatibility under a fixed structural scaffold.[24–26] These frameworks provide likelihood-based proxies for stability changes by evaluating the probability of a specific residue within the wild-type structural context, thereby bypassing the need for explicit mutant structure construction.[27–31] A prominent example is ThermoMPNN, which adapts the ProteinMPNN architecture to predict $\Delta\Delta G$ based on these log-

[a]State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China. E-mail: xiongyi@sjtu.edu.cn

[b]Zhangjiang Institute for Advanced Study, Shanghai Jiao Tong University, Shanghai, China

[c]Shanghai Innovation Institute, Shanghai, China

likelihood scores.[27] Finally, a subset of methods explicitly uses mutant structures generated by tools such as AlphaFold series,[32,33] enabling direct comparison between wild-type and mutant conformations,[34–36] as exemplified by ThermoNet.[34]

Despite these advances, accurately modeling the stability change of complex genetic variation remains challenging, with distinct limitations arising from different classes of methods.[37,38] For physics-based approaches, predictive accuracy is often hindered by limited conformational sampling and biases in the underlying energy potentials.[39] Data-driven models face a different set of challenges. While many frameworks can process multi-site substitutions, they often rely on simplistic additivity assumptions that neglect epistatic interactions.[30] Moreover, indels, which fundamentally reshape the protein backbone and alter long-range interactions, remain largely unexplored.[40,41] Furthermore, a central challenge concerns how structural information is incorporated in structure-based deep learning models. Approaches that condition on a fixed structural scaffold—such as microenvironment encoding on a wild-type backbone and inverse folding under a static backbone—cannot represent the conformational plasticity induced by indels and other large-scale perturbations. By contrast, geometry-based pipelines that predict structural intermediates avoid requiring experimental structures, but typically introduce an intermediate representation bottleneck. These simplified geometric descriptors are often not learned adaptively for the stability change prediction task, failing to capture the fine-grained structural nuances essential for $\Delta\Delta G$ prediction. Conversely, approaches that explicitly generate mutant structures to model backbone changes incur prohibitive computational costs, limiting scalability. Finally, most deep learning approaches remain difficult to interpret, offering limited insight into why particular mutations confer stability and providing little guidance for active sequence optimization.

To address these limitations, we introduce UniStab, a unified deep learning framework capable of predicting stability changes across single-point, multi-point and indel mutations in an end-to-end manner. UniStab does not rely on pre-computed structural features or explicit structure generation. Instead, it extends the folding trunk of a protein structure prediction model with low-rank adaptation (LoRA),[42] thereby interrogating the model's latent structural reasoning. This design enables UniStab to capture thermodynamic perturbations arising from local backbone rearrangements and residue couplings directly from implicit representations, bypassing the trade-off between the inaccuracy of fixed-backbone approximations and the high cost of explicit modeling. UniStab achieves state-of-the-art performance, outperforming existing methods by 11% on multi-point mutations and 19% on indels. Beyond predictive accuracy, UniStab offers structural interpretability through attention-based analysis of residue interactions and facilitates practical protein engineering when coupled with a beam-search optimization strategy. Collectively, these capabilities—unified prediction, mechanistic interpretability, and generative design—establish UniStab as a versatile foundation for protein stability modeling and rational engineering.

2 Results

2.1 Overview of UniStab

An overview of the UniStab architecture is shown in Fig. 1. UniStab is designed as a unified framework for predicting stability changes across single-point mutations, multi-point mutations, and indels, thereby extending beyond the mutation regimes typically covered by existing stability change predictors.

As illustrated in Fig. 1a, UniStab takes the wild-type and mutant sequences as input and first encodes each sequence using ESM2-3B[19] to obtain residue-level sequence embeddings. Pair representations are initialized at the input to the folding trunk and jointly refined with the sequence representations through repeated trunk blocks. The resulting latent representations are further processed by a compact structure module, weighted attention pooling, and an MLP regressor to predict the thermodynamic stability change ($\Delta\Delta G$). By processing wild-type and mutant sequences within a shared architecture, UniStab learns mutation-aware representations for end-to-end stability change prediction across diverse mutation types.

A central idea of UniStab is to exploit implicit structural information from a pre-trained folding model without explicitly generating mutant 3D structures. Protein structure prediction models such as ESMFold[19] and AlphaFold2 (ref. 32) produce atomic coordinates by decoding latent representations into structure. Instead of performing explicit coordinate decoding, UniStab directly repurposes the folding-trunk representations as structurally informed features for stability change prediction. Because these trunk representations are optimized in the context of structure prediction, they encode geometric constraints and residue–residue dependencies that are informative for mutation-induced stability changes, while avoiding the computational cost of explicit mutant-structure generation.

The folding trunk used in UniStab is schematically shown in Fig. 1b. Starting from ESM2-derived sequence embeddings and initialized pair representations, the trunk jointly refines sequence-level and pairwise representations through repeated folding blocks. These coupled updates enable information exchange between residue-level features and pairwise interaction features, allowing UniStab to capture both local sequence context and long-range residue dependencies. This design is important for modeling non-additive effects in multi-point mutations and structural perturbations associated with indels.

To adapt the pre-trained folding representations efficiently, UniStab incorporates low-rank adaptation (LoRA; Fig. 1c). The original trunk weights are kept frozen, while lightweight trainable low-rank updates are introduced to refine the latent representations for stability change prediction. This parameter-efficient strategy enables task-specific adaptation without retraining the full folding model.

After trunk processing, the resulting latent representations are integrated by a compact structure module (Fig. 1d), which combines sequence and pair features through attention-based interactions and residual updates. The residue-level representations are then pooled into protein-level embeddings. Finally,

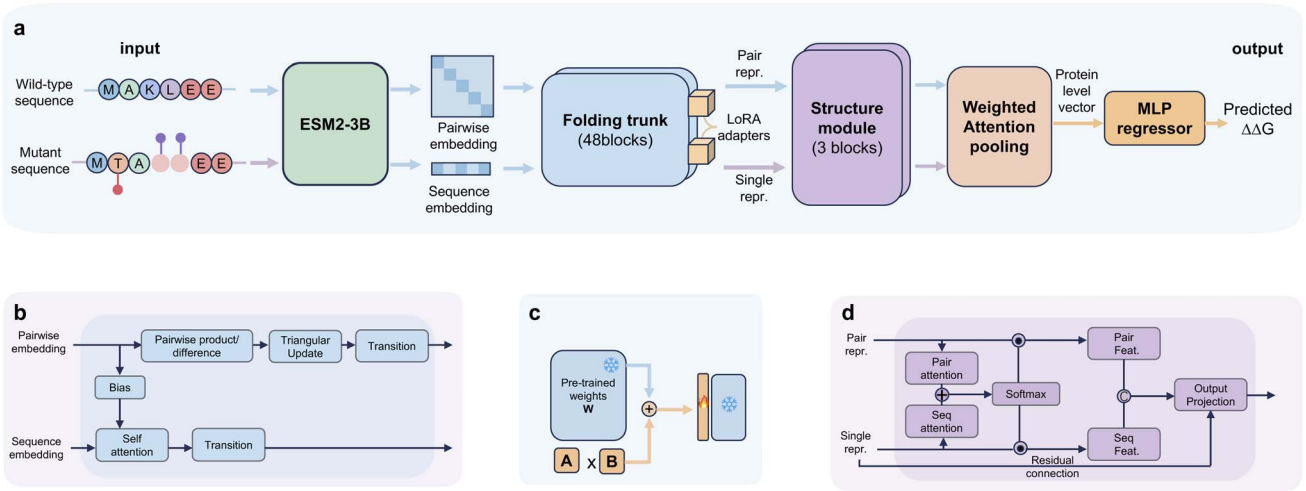

Fig. 1 Overview of UniStab. (a) Overall architecture of UniStab. Wild-type and mutant sequences are first embedded by ESM2-3B to obtain sequence-level representations. Pair representations are initialized at the input to the folding trunk and are jointly refined with the sequence representations through trunk blocks. The resulting latent representations are passed to a compact structure module, followed by weighted attention pooling and an MLP regressor for prediction of stability change (ΔΔ*G*). (b) Schematic of a folding trunk block, which jointly updates sequence and pair representations through coupled attention and pairwise operations. (c) Low-rank adaptation strategy used to efficiently adapt the pre-trained folding trunk. The original pre-trained weights are frozen, while trainable low-rank updates are introduced for task-specific refinement. (d) Schematic of the structure module, which integrates sequence and pair representations to generate residue-level features for stability change prediction.

the pooled mutant and wild-type protein-level embeddings are compared to form a differential representation, which is mapped by an MLP regressor to the predicted ΔΔ*G*.

Together, UniStab combines unified mutation coverage, implicit structural encoding, and parameter-efficient adaptation to provide a general framework for stability change prediction across diverse mutation types.

2.2 Benchmark datasets and comparative evaluation

To evaluate UniStab's performance across diverse mutational regimes, we utilized the high-throughput experimental data generated *via* cDNA display proteolysis. This dataset, known as MegaScale,[43] provides over 776 000 high-quality measurements, making it the most extensive resource for modeling protein thermodynamics. After quality filtering, we retained 470 995 mutations. We partitioned proteins by 25% sequence identity clustering to prevent homologous leakage across training, validation, and test sets, yielding 373 325 training, 43 070 validation, and 54 600 test samples (SI Table S1).

We assessed UniStab's performance across three primary categories within the MegaScale test set: single-point mutations, double mutations, and indels. To examine the model's generalization capability on complex, higher-order variants, we further employed the PTMUL dataset[44] as an independent benchmark. We compared UniStab against a broad spectrum of methods, including current state-of-the-art deep learning models tailored to specific mutation types (*e.g.*, ThermoMPNN-D,[30] SPURS,[28] and MutateEverything[23]), widely used physics-based approaches (*e.g.*, FoldX,[14] Rosetta[15]), and other representative predictors such as ESM2-3B[19] and DDGun.[44] Detailed descriptions of these datasets and method configurations are provided in the Methods section.

Since UniStab predicts continuous ΔΔ*G* values, we evaluated its performance from both regression and classification perspectives. For regression, we report Spearman's rank correlation (SCC), Pearson correlation (PCC), root mean squared error (RMSE), and mean absolute error (MAE). For classification, mutations were binarized as stabilizing or destabilizing, and we report AUROC, AUPRC, F1 score, and Matthews correlation coefficient (MCC). Top-*k*% accuracy was used to measure the ability to identify the most stabilizing variants. Full definitions are provided in Methods.

2.3 UniStab achieves strong and balanced performance on the MegaScale test set

As shown in Fig. 2, UniStab performed best on indels (SCC = 0.78, PCC = 0.78), followed by single-point mutations (SCC = 0.67, PCC = 0.75), while double mutations proved more challenging (SCC = 0.60, PCC = 0.62). Overall, these results indicate high predictive performance across mutation types, with particularly robust generalization to indels.

UniStab achieved balanced classification performance across mutation types despite the pronounced class imbalance in the MegaScale test set (Fig. 3a). As shown in Fig. 3b–d, single mutations yielded an F1 score of 0.66 and AUPRC of 0.53; indels achieved $F1 = 0.46$ and AUPRC = 0.34. Double mutations posed the greatest challenge ($F1 = 0.404$, AUPRC = 0.111), consistent with the extreme imbalance in this category where over 95% of variants are destabilizing.

We further assessed ranking performance, which is particularly relevant for protein engineering applications (Fig. 3e and f). UniStab achieved strong ranking on single mutations and indels, with the model's top-1 prediction falling within the top 3% and 8% of the ground-truth stability ranking, respectively.

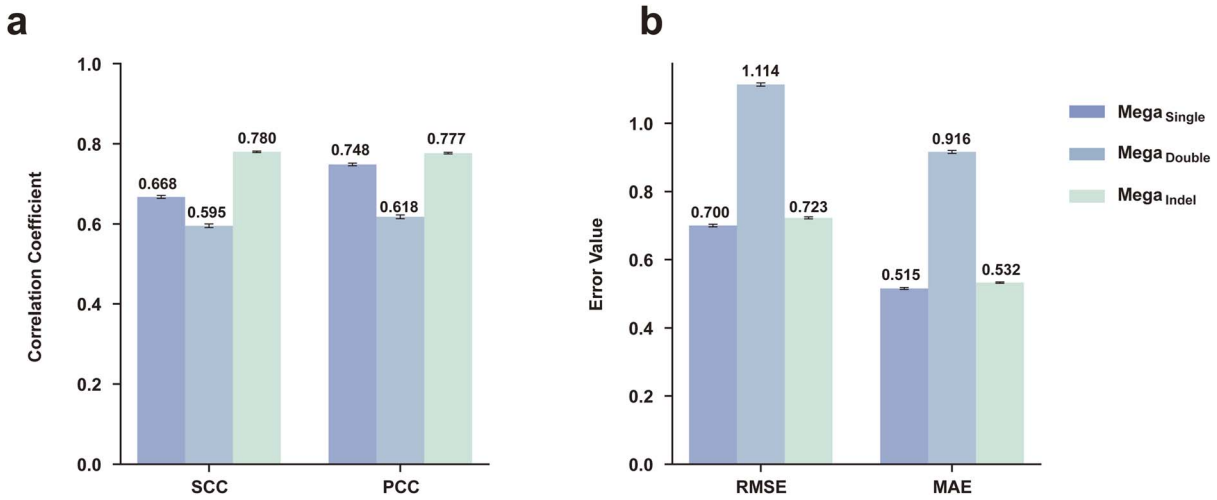

Fig. 2 Performance on the MegaScale test set. (a) Correlation coefficients (SCC, PCC) for UniStab across $Mega_{Single}$, $Mega_{Double}$, and $Mega_{Indel}$. (b) Error metrics (RMSE, MAE) on the same $Mega_{Single/Double/Indel}$ subsets; lower is better. Error bars indicate bootstrap standard deviations of the evaluation metrics, estimated from 1000 resampled test sets for each MegaScale subset.

In contrast, double mutations showed weaker ranking performance (87th percentile), consistent with their overall prediction difficulty. Top-5% accuracy was highest for indels (0.33), followed by single mutations (0.23) and double mutations (0.18).

We next examined how structural context affects prediction accuracy by partitioning double mutations based on solvent-accessible surface area (SASA) into core–core, core–surface, and surface–surface categories (see Methods for details). As shown in Fig. 3h–j, UniStab achieved comparable performance on surface–surface (PCC = 0.662, SCC = 0.655) and core–surface mutations (PCC = 0.674, SCC = 0.618), but substantially lower accuracy on core–core mutations (PCC = 0.426, SCC = 0.399). This pattern suggests that the model's implicit structural representations capture solvent-exposed features more effectively than densely packed hydrophobic cores.

Finally, we stratified double mutations by Cα–Cα distance into close-range (<5 Å), medium-range (5–10 Å), and distant (>10 Å) categories (Fig. 3k). As shown in Fig. 3l, UniStab achieved the strongest correlation on close-range pairs (PCC = 0.837, SCC = 0.815), with performance decreasing substantially for medium-range (PCC = 0.595, SCC = 0.570) and distant pairs (PCC = 0.531, SCC = 0.519). This distance-dependent trend suggests that the model more effectively captures local cooperative effects between spatially proximal mutations.

2.4 UniStab outperforms existing stability change predictors across benchmark datasets

We benchmarked UniStab against competitive methods across four datasets (Fig. 4). On $Mega_{Double}$, UniStab achieved the highest correlation (SCC = 0.595), surpassing SPURS (0.584) and ThermoMPNN-D (0.557). On $Mega_{Indel}$, UniStab (SCC = 0.780) markedly outperformed ESM2-3B (0.589), underscoring its robustness for indels. On $PTMUL_{Double}$, ThermoMPNN-D held a slight lead (SCC = 0.566 *vs.* 0.555), likely reflecting its specialization for pairwise substitutions. On $PTMUL_{Multi}$ (≥3 mutations), UniStab (SCC = 0.713) outperformed MutateEverything (0.600), demonstrating its advantage in modeling higher-order variants. Overall, UniStab ranks first on three of four benchmarks, with particular strength in indel and multi-mutation scenarios.

We next focused on the PTMUL datasets for in-depth comparison. On $PTMUL_{Double}$, the ground-truth $\Delta\Delta G$ distribution is more balanced between negative and positive values than Mega (Fig. 5a). The predicted $\Delta\Delta G$ distributions for UniStab, ThermoMPNN-D, and SPURS are shown in Fig. 5b, and the dataset's Cα–Cα distance distribution in Fig. 5c.

We first report binary classification performance in terms of AUROC (Fig. 5d): UniStab achieved the highest AUROC (0.792), outperforming ThermoMPNN-D (0.779) and SPURS (0.568). AUPRC results (Fig. 5e) show a consistent trend—UniStab 0.599, ThermoMPNN-D 0.528, SPURS 0.313—indicating stronger ability to distinguish stabilizing ($\Delta\Delta G < 0$) from destabilizing ($\Delta\Delta G > 0$) variants on this dataset.

To assess how spatial proximity between mutated residues influences performance, we stratified $PTMUL_{Double}$ variants by Cα–Cα distance into close (<5 Å), medium (5–10 Å), and distant (>10 Å) categories (Fig. 5f). All three methods degrade from close to distant bins, suggesting that long-range interactions are generally harder to capture. Across all distance categories, ThermoMPNN-D outperformed UniStab, likely due to its explicit incorporation of inter-residue distance features in the model architecture—a design supported by ablation studies in the original publication.[30]

We further analyzed performance across $\Delta\Delta G$ ranges (Fig. 5g). In the extreme stabilizing regime ($\Delta\Delta G < -2$), all methods exhibited negative SCC values. This likely reflects a train–test distribution mismatch: models were trained on Mega, where ground-truth $\Delta\Delta G$ values are mostly confined to [−2, + 5], leading to regression saturation beyond this interval. PTMUL, by contrast, includes variants as stabilizing as ∼−6, explaining the loss of correlation for $\Delta\Delta G < -2$. Within the sampled range, performance diverged across models. In the

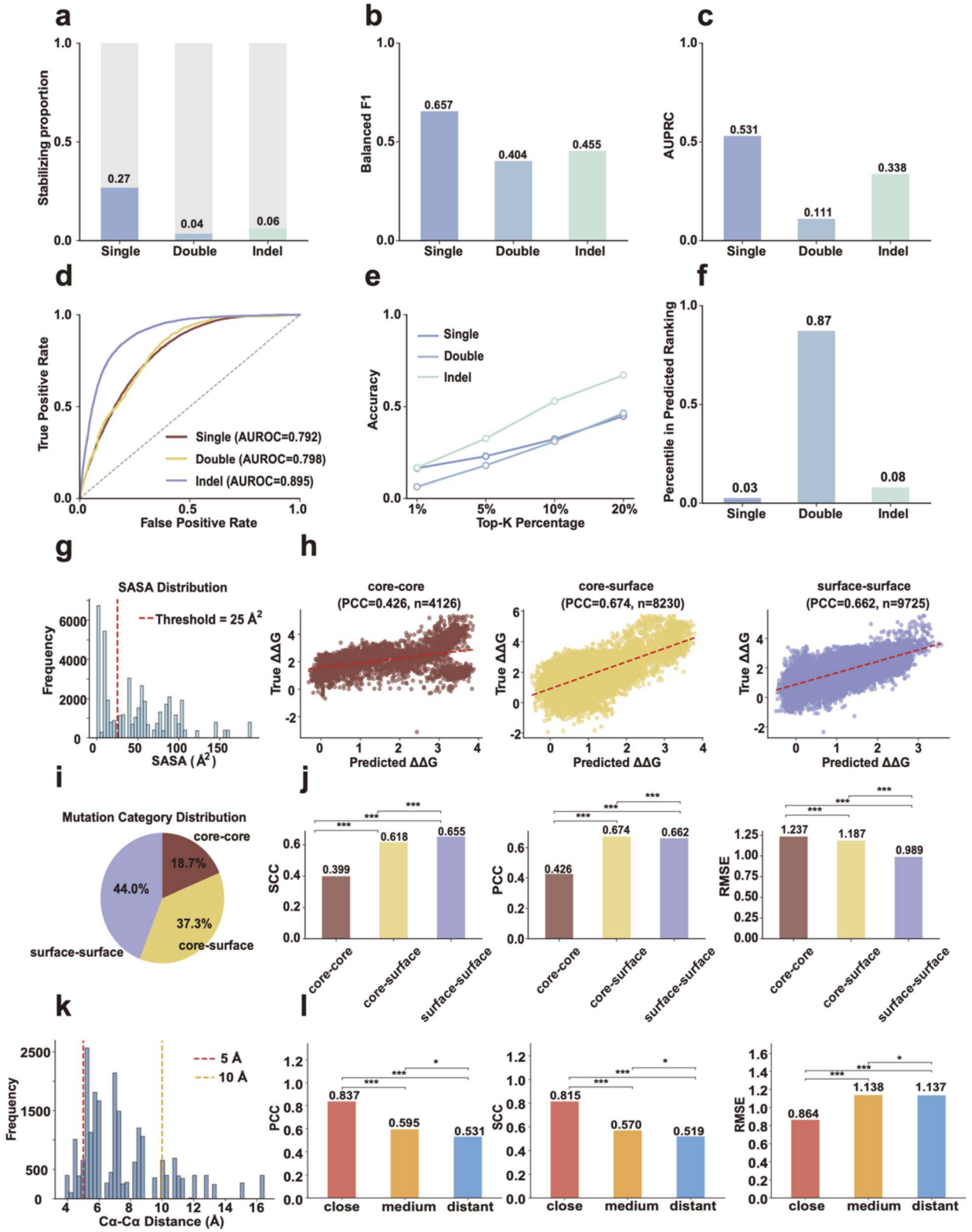

Fig. 3 Detailed performance analysis on the MegaScale test set. (a) Proportion of stabilizing *versus* destabilizing mutations by mutation type (single, double, indel). (b–d) Balanced classification metrics for UniStab on each mutation type: F1 score, area under the precision–recall curve (AUPRC), and area under the receiver operating characteristic curve (AUROC). (e) Top-*k*% accuracy ($k = 1, 5, 10, 20$; *k* denotes percentage of the test set). (f) Percentile position of the model's top-1 prediction in the ground-truth stability ranking (lower is better). (g) Distribution of per-residue solvent-accessible surface area (SASA) with threshold at 25 Å^2 (residues with SASA ≥25 Å^2 classified as surface). (h) Predicted *versus* experimental $\Delta\Delta G$ by SASA context (core–core, core–surface, surface–surface). (i) Proportion of double mutations in each SASA context. (j) Performance by SASA context (SCC, PCC, RMSE). (k) Distribution of inter-residue Cα–Cα distances for double mutations, with cutoffs at 5 and 10 Å. (l) Performance by distance bin (SCC, PCC, RMSE).

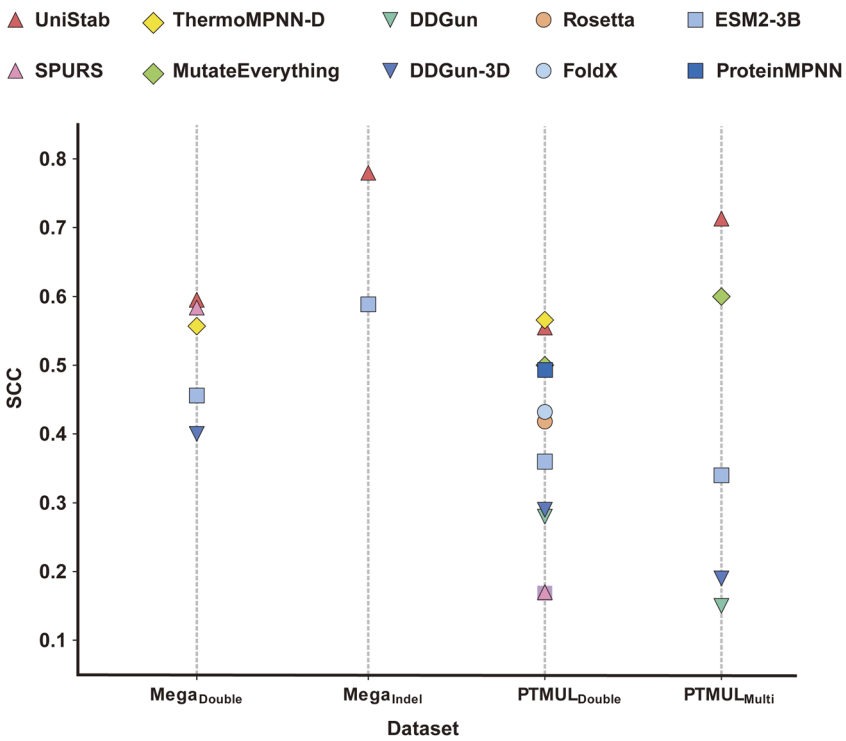

Fig. 4 Comparison with competitive methods. SCC of UniStab and competitive methods across four benchmarks: $Mega_{Double}$, $Mega_{Indel}$, $PTMUL_{Double}$, and $PTMUL_{Multi}$. UniStab (red triangle) exhibits superior generalization, particularly in challenging indel and multi-point mutation scenarios.

moderately stabilizing region ($-2 \leq \Delta\Delta G < 0$), both ThermoMPNN-D and SPURS outperformed UniStab. One possible explanation is that these inverse-folding models estimate sequence likelihoods conditioned on backbone geometry, which may implicitly link amino acid probability with structural tolerance and energetic stability. Such a formulation could make them more sensitive to subtle stabilizing substitutions that preserve local packing. By contrast, UniStab, which relies on implicit structural representations, performed best for destabilizing mutations ($0 \leq \Delta\Delta G \leq 4$), potentially reflecting its capacity to detect energetic perturbations that disrupt local structure. These observations collectively suggest that explicit and implicit structure modeling paradigms may emphasize different energetic regimes.

We then turned to $PTMUL_{Multi}$ to examine multi-mutation scenarios. The ground-truth $\Delta\Delta G$ distribution is shown in Fig. 5h, and the frequency of mutation counts (3–10) in Fig. 5i. The confusion matrix (Fig. 5j) summarizes class-wise prediction errors, whereas the SCC–mutation-count curve (Fig. 5k) reveals an overall upward trend—from 0.71 at three mutations to 0.89 at nine—with minor fluctuations. Such a pattern likely reflects that larger deviations from the wild type amplify structure-dependent energetic signals captured by UniStab's implicit representations. Finally, Fig. 5l compares UniStab and MutateEverything on the $PTMUL_{Multi}$ dataset. UniStab achieves a higher MCC (0.716 *vs.* 0.580) and AUROC (0.884 *vs.* 0.860), indicating that it ranks variants better across thresholds and makes more consistent binary decisions. Together, these findings highlight UniStab's robustness and its capacity to generalize effectively to complex, high-order mutational landscapes where structure-based signals become increasingly informative.

2.5 Ablation analyses identify critical architectural and fine-tuning components

We conducted ablation studies on three design factors: the number of blocks in the structure module, the pooling strategy, and the LoRA configuration (Fig. 6; SI Table S9).

Across single, double, and indel mutation tasks, the structure module proved essential: removing it reduced SCC from 0.668 to 0.466 on $Mega_{Single}$ and from 0.595 to 0.400 on $Mega_{Double}$. Increasing the number of blocks from one to three consistently improved correlations and reduced errors, whereas adding a fourth block provided no further gains, suggesting that performance saturates at three blocks.

Regarding pooling, the full configuration—three structure blocks with weighted attention pooling and LoRA (rank $= 8$, $\alpha = 16$)—performed best overall. Mean pooling achieved slightly higher SCC for single mutations (0.677 *vs.* 0.668) but yielded higher errors; for double mutations it reduced errors but slightly lowered correlations. For indels, the full configuration achieved the best results across all metrics (SCC $= 0.780$), with mean pooling ranking at the second position (0.704).

The LoRA ablation further confirmed the importance of capacity: reducing the rank and α from (8, 16) to (4, 8) decreased SCC from 0.780 to 0.629 on indels and from 0.595 to 0.503 on double mutations, indicating that higher LoRA capacity better adapts the folding trunk's implicit structural representations.

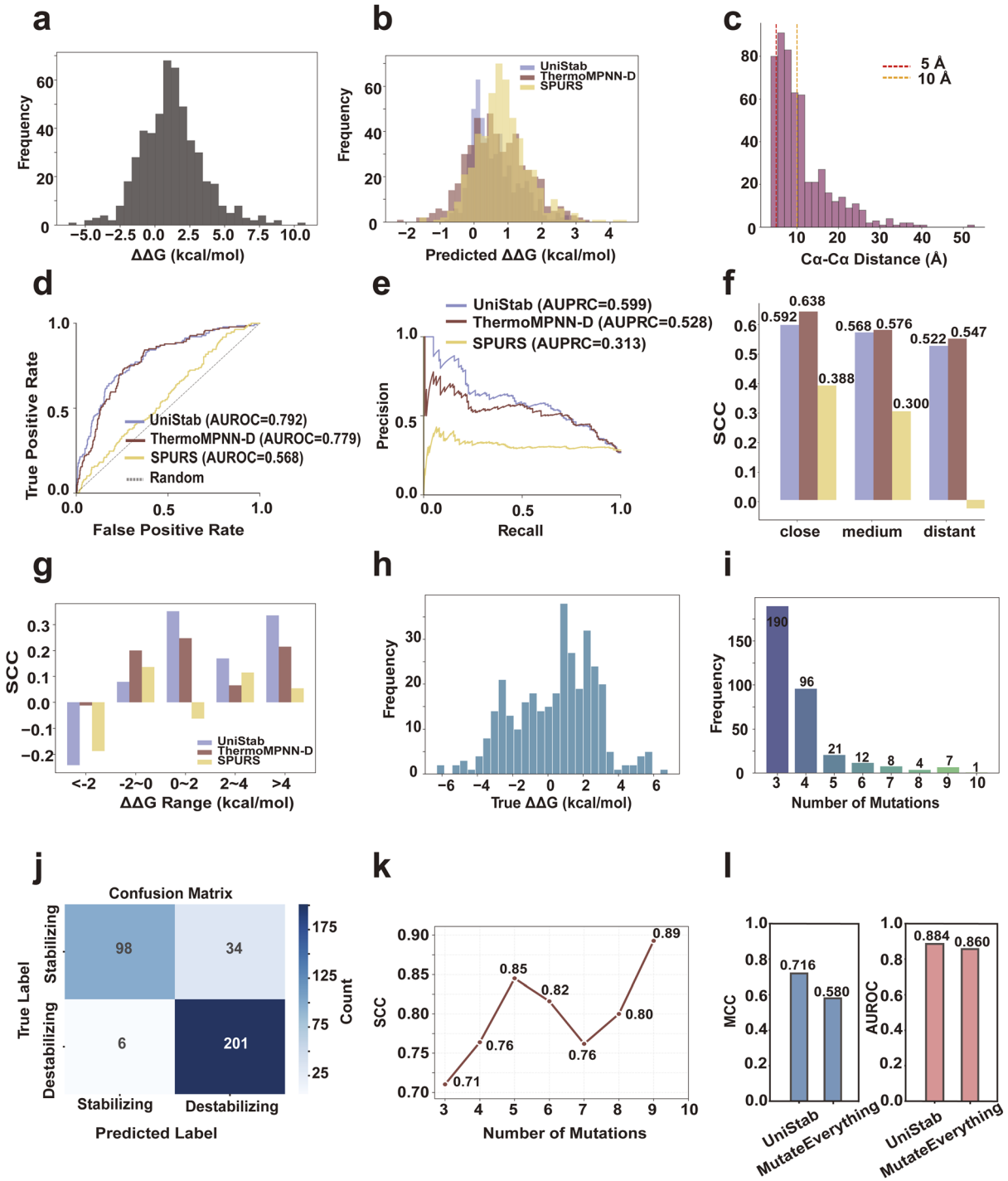

Fig. 5 Performance comparison on the PTMUL datasets. (a) Ground-truth $\Delta\Delta G$ distribution for $PTMUL_{Double}$. (b) Predicted $\Delta\Delta G$ distributions for UniStab, ThermoMPNN-D, and SPURS. (c) Cα–Cα distance distribution in $PTMUL_{Double}$. (d) Binary classification AUROC for the three methods. (e) Binary classification AUPRC. (f) Performance across distance bins: close (<5 Å), medium (5–10 Å), distant (>10 Å). (g) SCC across $\Delta\Delta G$ ranges. (h) Ground-truth $\Delta\Delta G$ distribution for $PTMUL_{Multi}$. (i) Frequency distribution of mutation counts (3–10). (j) Confusion matrix for UniStab on $PTMUL_{Multi}$. (k) SCC *versus* mutation count for UniStab. (l) MCC and AUROC comparison between UniStab and MutateEverything on $PTMUL_{Multi}$.

Overall, combining a three-block structure module, weighted attention pooling, and high-capacity LoRA yields the best balance between correlation and error, achieving the aggregate performance—particularly on indel mutations.

2.6 Feature distance correlates with prediction accuracy and captures local structural perturbations

We next investigated why SCC performance increases with mutation count (Fig. 7). We hypothesized that more mutations increase the representational distance between mutant and

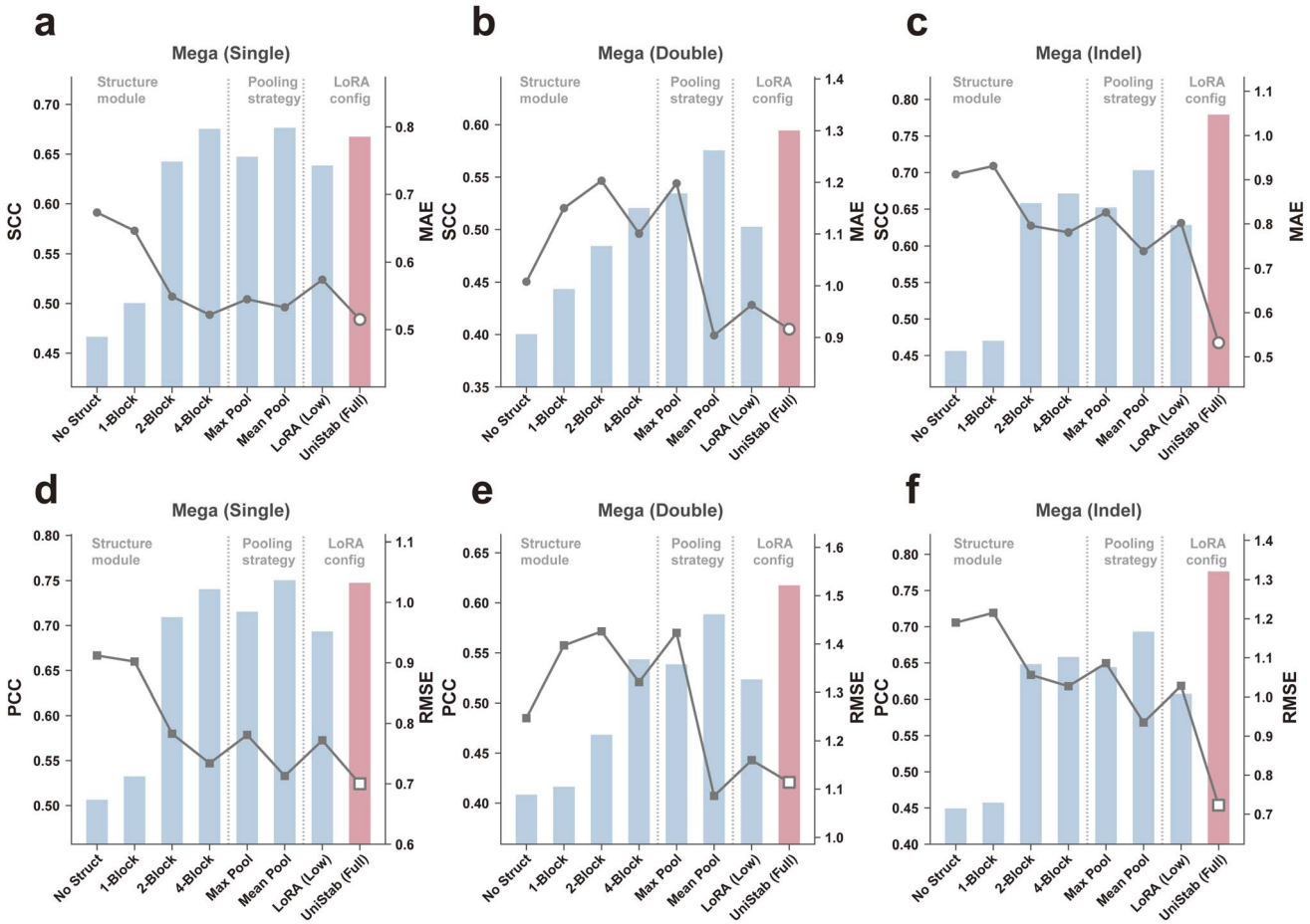

Fig. 6 Ablation analysis of key architectural components and training strategies. The impact of the structure module depth, pooling mechanism, and LoRA capacity on model performance is evaluated across MegaScale single-point (a, d), double-point (b, e), and indel (c, f) mutation tasks. (a–c) Performance metrics SCC (bars, left axis) and MAE (lines, right axis). (d–f) Performance metrics PCC (bars, left axis) and RMSE (lines, right axis). In all panels, blue bars represent intermediate configurations, while red bars highlight the final UniStab model (structure module: 3-block; pooling: weighted attention; LoRA: high-rank). Configurations compared include: (1) structure module: varying the number of stacked blocks (no struct, 1-block, 2-block, 4-block) *versus* the optimal 3-Block design. (2) Pooling strategy: comparing standard Max and Mean pooling against the learnable weighted attention pooling. (3) LoRA config: contrasting a low-rank setting (rank = 4, alpha = 8) with the high-capacity setting (rank = 8, alpha = 16). Across all mutation types, the full UniStab configuration consistently achieves an optimal balance between high correlation and low error.

wild-type proteins, enhancing their separability in feature space. Consistent with this hypothesis, normalized feature-space distance rose with mutation number in the PTMUL dataset (Fig. 7a). We grouped samples into ten bins by feature distance (Fig. 7b). Across all samples, UniStab achieved PCC = 0.491 and SCC = 0.587 (Fig. 7c), with both metrics improving steadily from Bin 1 to Bin 10 (Fig. 7d). Mean feature distance per bin was positively correlated with SCC ($r = 0.751$, $p = 0.012$; Fig. 7e), confirming that greater representational separation is associated with higher prediction accuracy.

To validate that UniStab's feature distance reflects genuine structural perturbations, we employed Chai-1,[45] an independent structure prediction model, to generate mutant structures and calculated local RMSD between predicted mutant and wild-type structures. The Chai-1 predictions showed high confidence across all samples (mean pLDDT = 94.4; Fig. 7f), with pLDDT remaining stable across different mutation counts (Fig. 7g), confirming reliable structure quality for this analysis. We focused on $\text{PTMUL}_{\text{Multi}}$ (≥3 mutations), as double mutations showed negligible correlation—consistent with our earlier observation that feature distance increases with mutation count (Fig. 7a). As shown in Fig. 7h, UniStab's feature distance exhibited a weak but positive correlation with Chai-1-derived local RMSD at both 4 Å (SCC = 0.103) and 8 Å (SCC = 0.209). Notably, UniStab consistently outperformed the sequence-only ESM2-3B baseline (SCC = 0.035 and 0.029, respectively), indicating that the structure module contributes to capturing local conformational changes. These results suggest that as mutation count increases, the induced structural perturbations become large enough for UniStab's representations to detect.

We also compared additive and epistatic approaches for multi-mutation prediction (Fig. 7i). UniStab-additive sums single-mutant predictions, while UniStab-epistatic directly predicts multi-mutant effects. Both approaches achieved comparable SCC across mutation counts (ranging from 0.69 to 0.89), with epistatic slightly outperforming additive in most

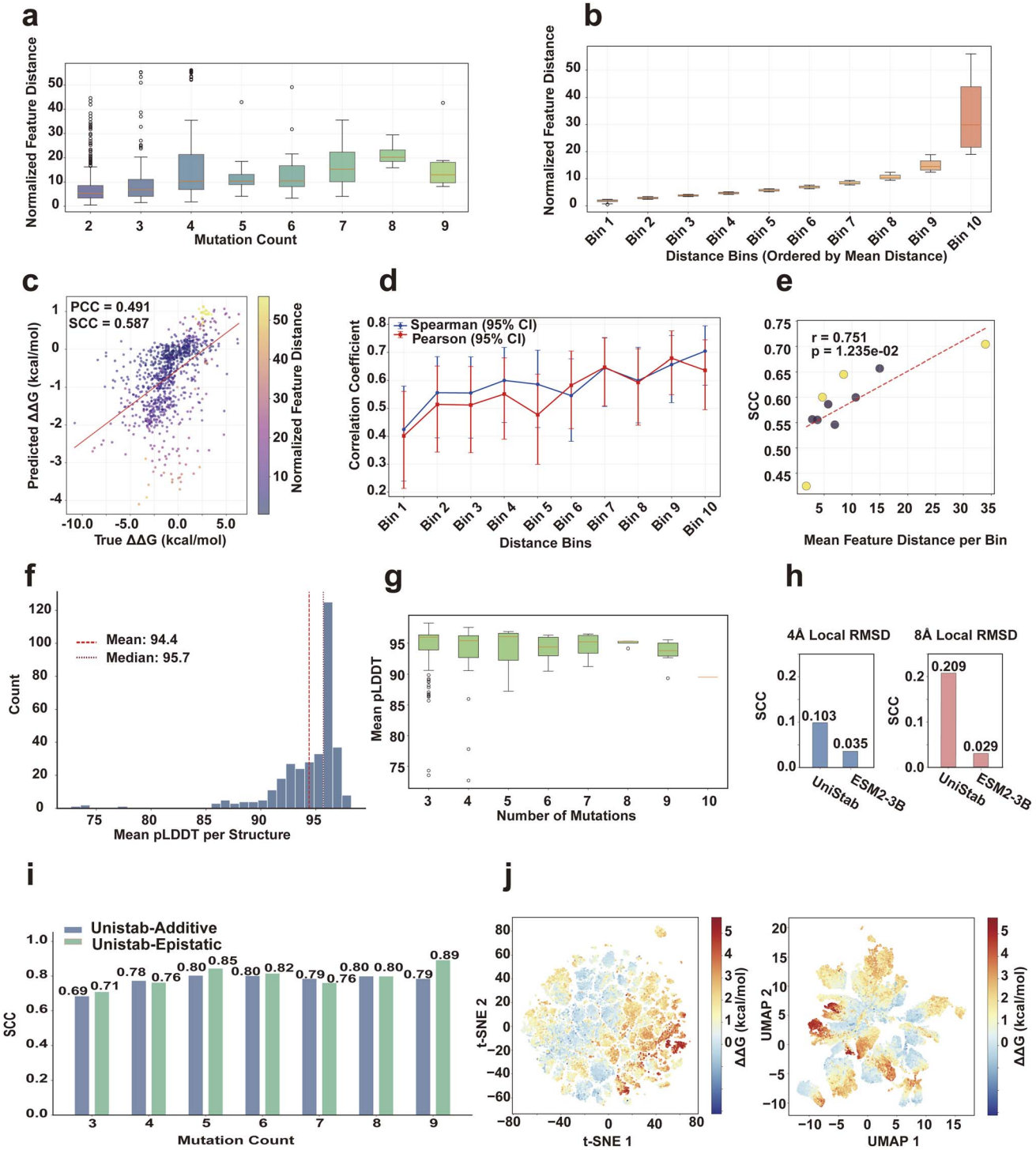

Fig. 7 Feature distance predicts performance and reflects local structural changes. (a) Normalized feature distance *versus* mutation count in PTMUL (double + multi). (b) Ten bins ordered by increasing feature distance (Bin 1–Bin 10). (c) Predicted *versus* experimental $\Delta\Delta G$ on PTMUL. (d and e) Performance increases across feature-distance bins; SCC correlates positively with mean feature distance. (f) pLDDT distribution of Chai-1-predicted mutant structures. (g) Mean pLDDT of Chai-1-predicted mutant structures across different mutation counts. (h) Feature distance correlates with local RMSD calculated within 4 Å and 8 Å radii around mutation sites and outperforms ESM2-3B as a local-change indicator. (i) UniStab-epistatic (direct multi-mutant prediction) slightly outperforms UniStab-additive (sum of single-mutant predictions) across mutation counts. (j) t-SNE/UMAP of difference features (mutant minus wild type) colored by experimental $\Delta\Delta G$, showing smooth transitions along the embedding. Definitions of feature distance and local RMSD are provided in Methods.

cases (*e.g.*, 0.85 *vs.* 0.80 at 5 mutations, 0.89 *vs.* 0.79 at 9 mutations). The overall similarity suggests that most mutations combine near-additively,[46] while the epistatic model provides modest gains for variants with stronger non-additive effects.

Finally, we projected difference features using t-SNE and UMAP, coloring points by experimental $\Delta\Delta G$ (Fig. 7j). Both embeddings displayed smooth color gradients from stabilizing (blue) to destabilizing (red) variants, indicating that the learned representations capture a continuous mapping of the stability landscape.

2.7 Attention highlights mutation-centered interaction rewiring and aligns with local structural changes

We used two representative cases to examine whether UniStab's attention maps capture mutation-centered changes in residue interactions. In the 1RGG multi-mutation series, extending the wild type (WT) to an eight-site mutant (M8) remained stabilizing (experimental $\Delta\Delta G = -6.3$), while adding a ninth mutation, I70D, to M8 (yielding M9) reversed this trend, converting the protein to destabilizing (experimental $\Delta\Delta G = +3.6$). UniStab correctly captured this transition, predicting M8 as stabilizing (-0.93) and M9 as destabilizing (+1.625). The WT, M8, and M9 structures (Fig. 8a–c) showed no major global conformational changes, indicating that this stability switch arises from local interaction rewiring rather than large-scale structural rearrangements.

To probe the internal signals, we constructed attention-score difference maps by subtracting attention scores between pairs of states (WT *vs.* M8, WT *vs.* M9, M8 *vs.* M9; Fig. 8d–f). In the M9–WT and M9–M8 comparisons (Fig. 8e and f), attention concentrated strongly on T18 and Y81 in the vicinity of D70 (the ninth mutation, I70D), whereas the M8–WT comparison (Fig. 8d) did not show comparable focus. We define a residue–residue interaction as a minimum heavy-atom distance $\leq$3.5 Å. Under this criterion, WT exhibits clear interactions between I70 and Y81/T18 (Fig. 8g); in M8, the I70–T18 interaction disappears (Fig. 8h); in M9, D70 interacts with Y81/T18, but the interaction is noticeably weaker than in WT (Fig. 8i). The spatial pattern of attention differences coincides with these interaction changes. The lack of a strong signal in M8–WT likely reflects the model's attention strategy: it first focuses on the mutated sites and then highlights nearby residues whose interactions change markedly due to the mutation, as in the I $\rightarrow$ D substitution in M9.

We further validated this behavior on a second protein composed mainly of α-helices. Fig. 8j shows the WT and a triple mutant (S14A, H18A, N21A) structures and residue mapping; Fig. 8k presents attention-score distributions for WT (left) and the mutant (right), and their differences (center). In the original attention maps, deep bands near the diagonal indicate strong attention between helix-adjacent positions, consistent with the i, $i \pm 3/i \pm 4$ spatial proximity pattern in α-helices. In the difference map, the largest changes concentrate on N16, Y17, and E20; structural mapping shows these residues lie close to the three mutated sites, indicating that attention differences are sensitive to local packing and interaction changes.

Together, these cases show that UniStab's attention mechanism centers on mutated residues and can indicate rewiring of interactions with nearby residues. These signals align with spatially resolved interaction changes and become more pronounced when mutations induce substantial local perturbations, providing micro-level interpretability that complements the relationship between feature distance and predictive performance described earlier.

2.8 Beam-search-guided optimization and molecular dynamics validation of enhanced PG variants

To enable automated discovery of stability-enhancing protein variants, we combined UniStab with a beam-search optimization algorithm, forming the UniStab–Beam framework (Fig. 9a). We demonstrated its effectiveness on protein-glutaminase (PG) as an example.

The optimization rapidly converged, yielding the top three variants, each containing two substitutions: L1M + A129C, L1M + S133I, and L1M + S167C. UniStab directly predicts $\Delta\Delta G$, with $\Delta\Delta G < 0$ denoting increased stability. The three top-ranked variants had predicted $\Delta\Delta G$ values ranging from -1.926 to -1.832 kcal·mol^{-1}, indicating favorable stabilizing effects. For beam-search ranking, we used the internal optimization score $S_{\mathrm{stab}} = -\Delta\Delta G$, such that larger scores correspond to more stabilizing predictions. The structures of the wild-type protein and its three variants are shown in Fig. 9b.

To validate these computational predictions, we performed 100 ns molecular dynamics (MD) simulations for the wild type and the three optimized variants. Root-mean-square deviation (RMSD) and root-mean-square fluctuation (RMSF) analyses showed that the variants displayed lower global and residue-level flexibility than the wild type (Fig. 9c and d). Native-contact analysis further revealed that all variants maintained higher fractions of native contacts throughout the simulations (Fig. 9e). The wild type exhibited a gradual decline from 100% to 87.75%, whereas the variants stabilized at 89.80–90.56%, with L1M + A129C showing the most consistent preservation ($\sim$90.56%). This enhanced retention of native contacts likely reflects strengthened intramolecular interactions, consistent with the reduced backbone fluctuations observed in the RMSF profiles. Radius-of-gyration (Rg) analysis corroborated these findings (Fig. 9f). The wild type exhibited a gradual expansion from 15.99 Å to 16.06 Å, while the variants remained more compact, with mean Rg values of 15.77–15.79 Å.

Together, these complementary structural metrics consistently demonstrate that UniStab–Beam-optimized PG variants exhibit enhanced compactness and thermodynamic stability, validating the predictive reliability of the framework and underscoring its potential as a generalizable, data-efficient strategy for improving protein stability.

3 Discussion

In this study, we present UniStab, a unified framework for predicting protein stability changes from implicit structural representations learned by protein folding models. By adapting the

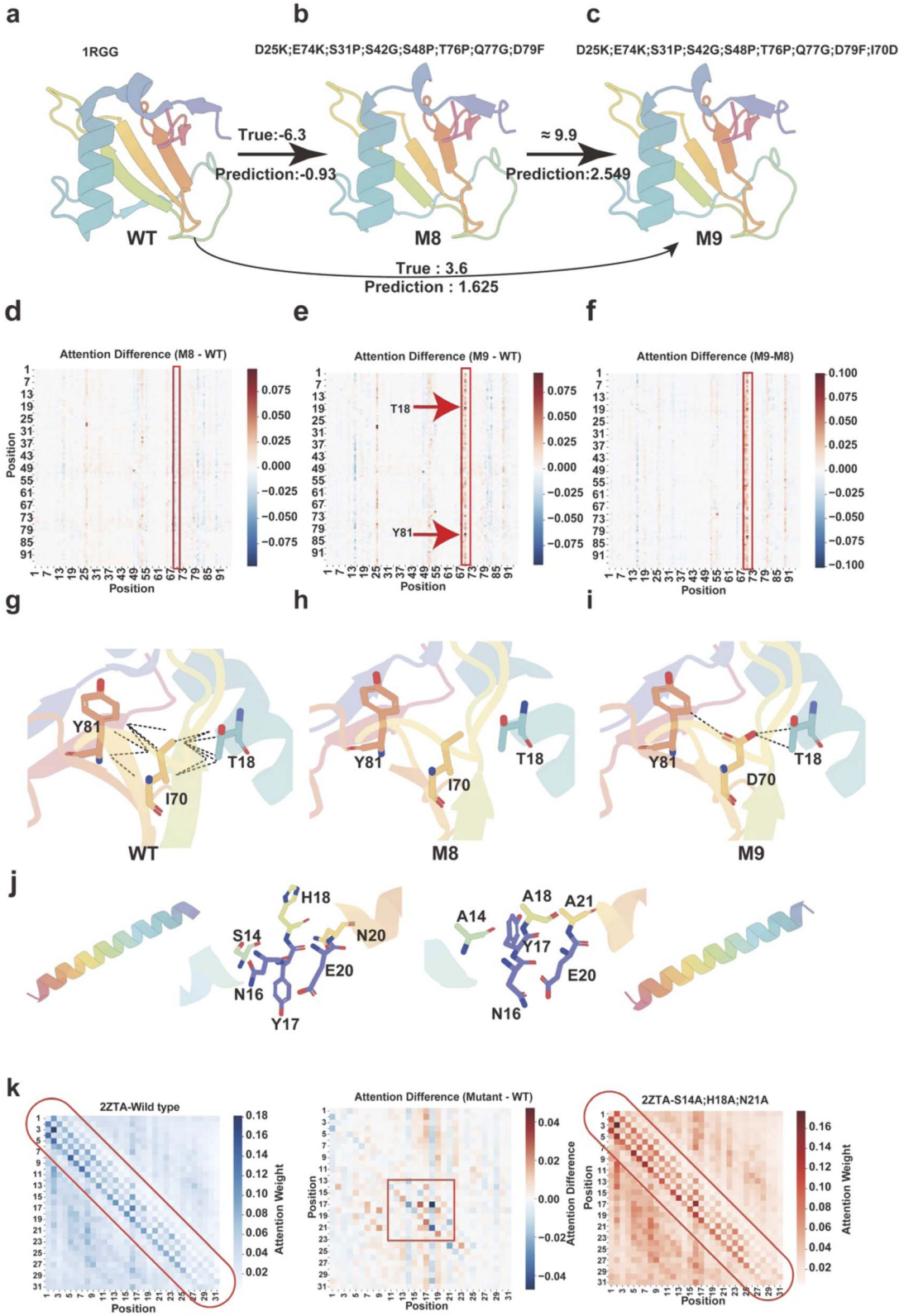

Fig. 8 Attention-based interpretation of mutation-induced interaction changes. (a–c) Predicted structures of 1RGG wild type (WT), the eight-site mutant M8, and the nine-site mutant M9 (M8 plus I70D), showing no major global conformational changes. (d–f) Attention-score difference maps for M8 *vs.* WT (d), M9 *vs.* WT (e), and M9 *vs.* M8 (f), highlighting changes near mutation sites. (g–i) Visualization of residue–residue interactions in WT (g), M8 (h), and M9 (i). Residue–residue interactions are defined as a minimum heavy-atom distance ≤3.5 Å; WT shows strong interactions between I70 and Y81/T18, which disappear in M8 and reappear in M9 but are weaker than in WT. (j) WT and triple mutant (S14A, H18A, N21A) structures of an α-helical protein, with mutated residues mapped. (k) Attention-score maps for WT (left) and mutant (right) and their differences (center); changes concentrate on residues near mutation sites, consistent with local packing modifications.

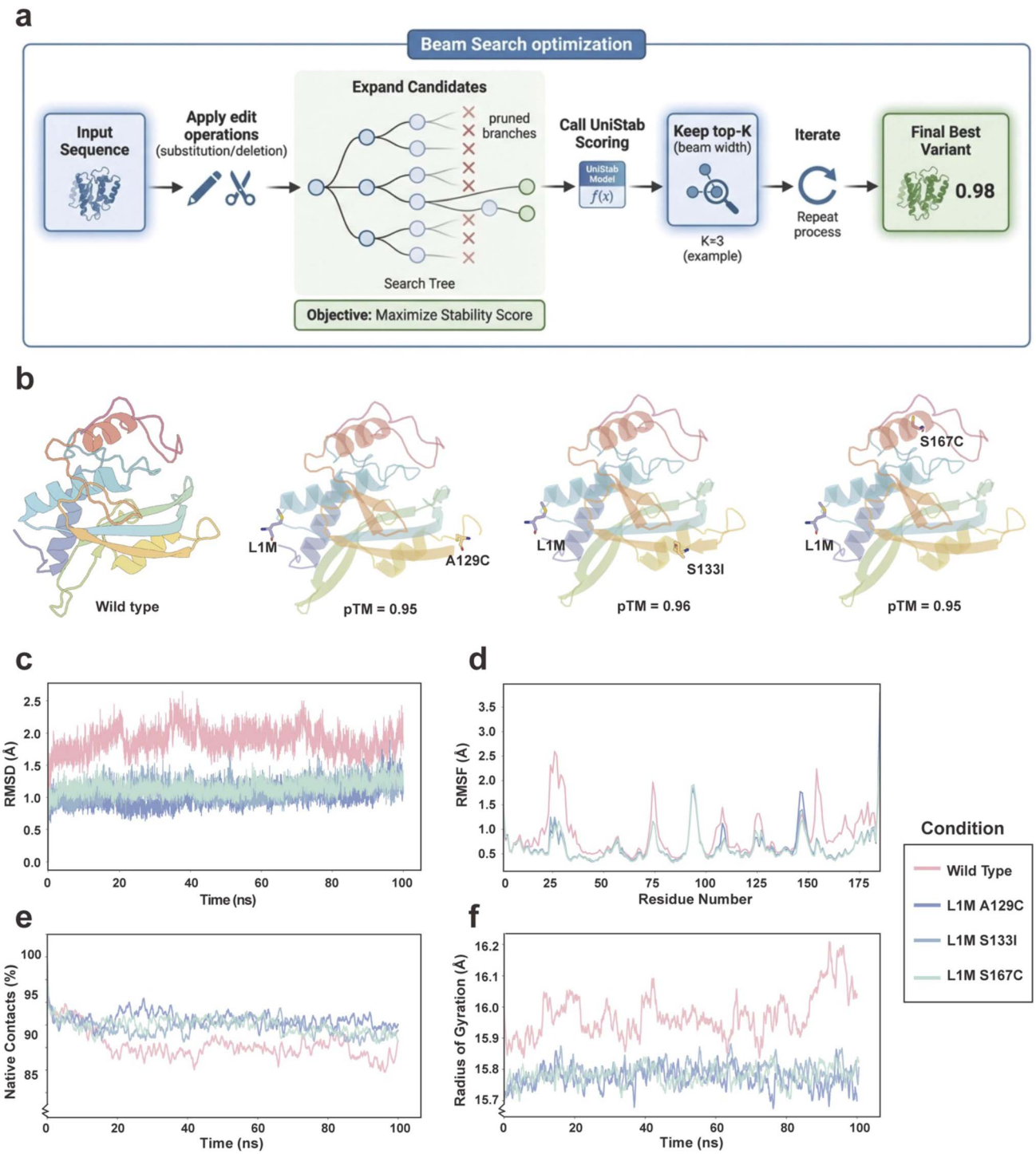

Fig. 9 Molecular dynamics validation of UniStab–Beam-optimized PG variants. Comparative analysis of structural dynamics for the wild-type protein and the three optimized variants (L1M + A129C, L1M + S133I, and L1M + S167C) during 100 ns MD simulations. (a) UniStab–Beam framework. (b) Structures of the wild-type protein and its three variants. (c) Root-mean-square deviation (RMSD) trajectories showing lower global fluctuations in all variants compared to the wild type. (d) Root-mean-square fluctuation (RMSF) profiles indicating reduced residue-level flexibility, particularly around the mutation sites. (e) Fraction of native contacts maintained during the simulations, demonstrating higher structural integrity in the variants relative to the wild type. (f) Radius of gyration (Rg) over time, showing more compact and stable conformations for the variants compared with the wild type.

folding trunk with low-rank adaptation, UniStab captures mutation-associated structural signals without requiring explicit mutant-structure generation for each candidate. This design reduces the computational cost associated with explicit structure-based prediction, while also avoiding the strict fixed-backbone assumption used in many inverse-folding-based approaches. As a result, UniStab provides a scalable framework for prediction of protein stability change that can be applied to single-point mutations, multi-point mutations, and indels.

Our analyses suggest that the predictive performance of UniStab is partly supported by mutation-associated structural signals encoded in the folding-trunk representations. Attention patterns were enriched around mutated residues and their contacting partners, and feature-distance analyses showed a relationship between wild-type/mutant embedding differences and local conformational changes. These observations indicate that UniStab captures local structural perturbations relevant to stability change prediction. Nevertheless, these analyses should be interpreted as supportive evidence rather than a complete mechanistic explanation of the model.

Beyond prediction, UniStab can also serve as a stability-guided objective for protein engineering. In a beam-search experiment on protein-glutaminase, UniStab prioritized stabilizing variants without exhaustive enumeration. In future work, UniStab could be integrated with reinforcement-learning[47,48] or generative[49–51] frameworks to guide sequence exploration toward candidates that balance stability with functional and evolutionary constraints.

In the present study, there remain at least two main limitations. First, although UniStab uses representations derived from a folding model, these latent features are still coarser than explicit three-dimensional structural models. This may limit its ability to capture long-range geometric consistency and nonlocal epistatic interactions, as reflected by the reduced performance on distant double mutations. Incorporating lightweight recycling or explicit residue–residue interaction constraints may improve the representation of long-range coupling while preserving scalability. Second, the generalizability of data-driven methods is still strongly influenced by the biased datasets such as MegaScale, including its protein coverage, dynamic range, and assay-specific characteristics. Therefore, generalization to underrepresented protein classes, experimental readouts, or mutation regimes should be interpreted with caution. Larger standardized datasets, broader external validation, and structure-informed synthetic data may help improve cross-protein generalization in future studies.

In summary, UniStab demonstrates that implicit structural representations from protein folding models can be adapted for scalable stability change prediction with structural interpretability. By unifying single-point mutations, multi-point mutations, and indels within one sequence-to-$\Delta\Delta G$ framework, UniStab offers a practical route to stability-aware protein engineering.

4 Methods

4.1 Datasets

The MegaScale dataset[43] comprises experimentally measured $\Delta\Delta G$ values for single, double, and indel mutations across a diverse set of proteins. We adopted the same clustering-based split strategy used in ThermoMPNN[27] and ThermoMPNN-D,[30] in which proteins are clustered at 25% sequence identity and splits are performed at the cluster level to ensure no homologous sequences appeared in both training and test sets. Unlike ThermoMPNN-D, which was trained exclusively on double mutations, UniStab was trained on all available mutation types in the dataset, including single-point, double-point, and indel mutations.

The PTMUL dataset[44] provides experimentally determined $\Delta\Delta G$ values for protein variants collected from multiple sources. The double-mutation subset ($\text{PTMUL}_{\text{Double}}$) used in this study comprises 536 double mutants across 83 proteins, curated by MutateEverything[23] and ThermoMPNN-D[30] following the quality-control and preprocessing protocols described in their studies. Since ThermoMPNN-D can only process double mutations, higher-order ($\geq$3) variants in PTMUL were excluded during their curation.

For multi-mutation ($\geq$3) analysis, we curated an extended $\text{PTMUL}_{\text{Multi}}$ dataset containing 339 samples from ProtDDG-Bench†, mapping each mutation set to its corresponding wild-type sequence. To ensure a fair head-to-head comparison with MutateEverything, we additionally evaluated the 310-sample subset used in their original paper, which is a subset of our curated 339-sample set. Unless otherwise noted, all multi-mutation analyses and interpretability experiments in this work were conducted on the complete 339-sample $\text{PTMUL}_{\text{Multi}}$ dataset.

4.2 Baseline methods

We primarily compared UniStab against three state-of-the-art deep learning methods designed for stability change prediction.

4.2.1 ThermoMPNN-D.[30] An extension of ThermoMPNN[27] tailored for double-mutant $\Delta\Delta G$ prediction. It adds directed-edge features between mutated sites to capture geometric context and uses a Siamese formulation to enforce invariance to mutation order. The model targets only double mutants and does not support single mutations, higher-order mutations ($\geq$3), or indels.

4.2.2 SPURS.[28] An adapter-based fusion of ProteinMPNN[24] geometry with ESM2 (ref. 19) sequence embeddings, enabling structure-informed single- and double-mutant prediction from wild-type inputs. It does not support higher-order mutations ($\geq$3) or indels.

4.2.3 MutateEverything.[23] A parallel decoding framework that predicts protein stability under multiple point mutations. We benchmark its AlphaFold-based, sequence-conditioned variant, which aggregates selected substitutions in latent space. It requires MSAs at inference, does not use mutant-specific structures, and cannot predict indels. For $\text{PTMUL}_{\text{Multi}}$ comparison, we used results reported in the original paper.

4.2.4 DDGun/DDGun3D.[52] An untrained method that predicts $\Delta\Delta G$ for single and multiple mutations using evolutionary information, with DDGun operating on sequence and DDGun3D additionally incorporating structural information.

A comprehensive summary of baseline capabilities, mutation type coverage, and whether results were re-evaluated or cited from original publications is provided in SI Table S6.

4.3 Evaluation metrics

The performance of UniStab was comprehensively evaluated using several metrics covering both regression and classification tasks.

4.3.1 Regression performance. For regression, the model's ability to predict $\Delta\Delta G$ values was quantified using PCC, SCC, RMSE, and MAE. PCC measures the linear correlation between predicted and experimental $\Delta\Delta G$ values, while SCC assesses the monotonic relationship between their ranks, providing robustness to non-linear effects and outliers.

RMSE evaluates the average magnitude of prediction errors, penalizing larger deviations more heavily, and is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\Delta\Delta G_{\mathrm{pred},i} - \Delta\Delta G_{\mathrm{exp},i}\right)^2} \tag{1}$$

MAE represents the mean absolute deviation between predictions and experiments, being less sensitive to large outliers:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\Delta\Delta G_{\mathrm{pred},i} - \Delta\Delta G_{\mathrm{exp},i}\right| \tag{2}$$

where n denotes the total number of mutations.

4.3.2 Classification performance. For classification analysis, mutations were categorized as stabilizing ($\Delta\Delta G_{\mathrm{exp}} < 0$) or destabilizing ($\Delta\Delta G_{\mathrm{exp}} > 0$). The model's discriminative capability was assessed using the area under the ROC curve (AUROC), area under the precision–recall curve (AUPRC), F1-score, and Matthews correlation coefficient (MCC). AUROC provides a threshold-independent measure of discriminative capability between classes, while AUPRC is particularly informative for imbalanced datasets. The F1-score and MCC offer threshold-dependent summaries of classification performance and are robust to class imbalance.

Additionally, the top-k% accuracy was computed for $k = 1$, 5, 10, and 20, measuring the precision among the top-ranked mutations predicted as most stabilizing. This metric directly evaluates the model's effectiveness in identifying promising candidates for protein engineering applications.

4.4 ESM2-3B embeddings and folding trunk

We initialized the sequence encoder and folding trunk from the pre-trained ESMFold v1 model and repurposed their latent representations for stability change prediction, without using the original coordinate-decoding structure module. Specifically, the ESM2-3B language model within ESMFold was used to generate residue-level sequence representations from single protein sequences, without requiring multiple sequence alignments (MSAs). For each input sequence of length L, we extracted the per-residue hidden representations returned by the ESM2 language model across all layers, including the input embedding layer, denoted as $h_i^{(l)} \in \mathbb{R}^{d_{\mathrm{model}}}$ for residue i at layer l, where $l = 0, \ldots, N_{\mathrm{layers}}$ and $d_{\mathrm{model}} = 2560$ for ESM2-3B.

Following the ESMFold architecture,[19] the layer-wise ESM2 representations were combined using learned scalar weights:

$$\tilde{h}_i = \sum_{l=0}^{N_{\mathrm{layers}}} \alpha_l h_i^{(l)}, \qquad \alpha_l = \frac{\exp(w_l)}{\sum_{m=0}^{N_{\mathrm{layers}}}\exp(w_m)} \tag{3}$$

where w_l are the pre-trained layer–combination parameters from ESMFold. In our implementation, the ESM2 language model and all original ESMFold parameters were frozen. The combined ESM2 representation was then projected to the folding-trunk sequence dimension and added to a learned amino acid embedding:

$$s_i^{(0)} = \mathrm{MLP}_{\mathrm{ESM}}\left(\tilde{h}_i\right) + E_{\mathrm{AA}}(a_i) \tag{4}$$

where a_i denotes the amino acid identity at residue i. This produces the initial sequence representation $S^{(0)} \in \mathbb{R}^{L\times d_{\mathrm{seq}}}$.

The pair representation was initialized at the input to the folding trunk. Specifically, the initial pair state was set to zero and augmented with the relative positional embedding from the ESMFold trunk:

$$P_{ij}^{(0)} = E_{\mathrm{relpos}}(r_i, r_j) \tag{5}$$

where r_i and r_j denote residue indices and E_{relpos} is the learnable pairwise positional embedding used by the folding trunk. Thus, the pair representation is not directly produced by ESM2-3B, but is initialized at the trunk input and subsequently refined together with the sequence representation.

The initial sequence and pair representations were then processed by the ESMFold folding trunk, consisting of 48 modified Evoformer-style blocks. Each trunk block jointly updates the sequence representation S and pair representation P through pair-biased self-attention, sequence-to-pair updates, triangular multiplicative updates, triangular attention, and transition layers. This iterative exchange allows the trunk to encode residue-level context and long-range pairwise dependencies in a structure-aware latent space.

To adapt the folding trunk for stability change prediction, we applied low-rank adaptation (LoRA) only to selected linear layers in the folding-trunk blocks, while keeping the original ESMFold parameters frozen. The output sequence and pair representations from the adapted folding trunk were used as latent structural features for stability change prediction, without invoking the original ESMFold coordinate-decoding structure module.

4.5 Structure-aware fusion module

We designed a compact structure-aware fusion module to integrate the sequence-level features $\boldsymbol{S} \in \mathbb{R}^{B\times L\times d_{\mathrm{seq}}}$ and pairwise features $\boldsymbol{P} \in \mathbb{R}^{B\times L\times L\times d_{\mathrm{pair}}}$ from the folding trunk. This module differs from the coordinate-decoding structure module in ESMFold: instead of generating 3D atomic coordinates, it fuses sequence and pair representations to produce residue-level features for stability change prediction. It comprises N_{blocks} stacked *StructureAwareFusion* layers, each using multi-head attention over sequence features modulated by pairwise biases.

In each block, sequence features are projected into query, key, and value tensors:

$$Q = \mathrm{proj}_q(\boldsymbol{S}) \tag{6}$$

$$K = \mathrm{proj}_k(\boldsymbol{S}) \tag{7}$$

$$V = \mathrm{proj}_v(\boldsymbol{S}) \tag{8}$$

where $Q, K, V \in \mathbb{R}^{B\times L\times H\times d_{\mathrm{hid}}}$.

Pairwise features are projected into head-specific attention biases $\boldsymbol{B}_{\mathrm{pair}} \in \mathbb{R}^{B\times L\times L\times H}$ and added to the scaled dot-product attention logits:

$$A_{bijh} = \mathrm{softmax}_j\left(\frac{\langle Q_{bih}, K_{bjh}\rangle}{\sqrt{d_{hid}}} + B_{\mathrm{pair},bijh}\right) \tag{9}$$

with padding positions masked when applicable.

The sequence context is computed as:

$$S_{\mathrm{ctx},bih} = \sum_{j=1}^{L} A_{bijh} V_{bjh} \tag{10}$$

The multi-head context is then concatenated across heads to form a residue-level sequence context vector.

In parallel, pairwise context for residue i is obtained by applying the attention weights to the pair features:

$$\boldsymbol{P}_{\mathrm{ctx},bi} = \sum_{j=1}^{L} \overline{\boldsymbol{A}}_{bij}\boldsymbol{P}_{bij}, \qquad \overline{\boldsymbol{A}}_{bij} = \frac{1}{H}\sum_{h=1}^{H} \boldsymbol{A}_{bijh} \tag{11}$$

The sequence and pairwise contexts are concatenated and projected back to the sequence dimension through a linear projection with residual connection and layer normalization:

$$\boldsymbol{S}' = \mathrm{LayerNorm}\left(\boldsymbol{S} + \mathrm{Proj}\left([\boldsymbol{S}_{\mathrm{ctx}}, \boldsymbol{P}_{\mathrm{ctx}}]\right)\right) \tag{12}$$

The output of each fusion block is used as the input to the next block. The final residue-level representation $\boldsymbol{S}^{(N_{\mathrm{blocks}})}$ is forwarded to the weighted attention pooling stage. Ablation experiments varying N_{blocks}, including removal of the structure-aware fusion module, were used to quantify the contribution of explicit pair–sequence fusion to the prediction performance.

4.6 Weighted attention pooling

To aggregate residue-level features from the structure-aware fusion module into a fixed-length variant-level representation, we employ a learnable global mask-weighted attention pooling layer. This design allows the model to focus on structurally or energetically important residues, rather than treating all positions equally.

Let $\boldsymbol{H} \in \mathbb{R}^{B\times L\times d}$ be the enhanced sequence features (B: batch size, L: sequence length, d: embedding dimension). A trainable weight vector $\boldsymbol{w} \in \mathbb{R}^d$ (optional bias $b \in \mathbb{R}$) computes a residue score:

$$e_i = \boldsymbol{w}^{\top}\boldsymbol{h}_i + b \tag{13}$$

where $\boldsymbol{h}_i$ is the i-th residue feature. A binary mask $\boldsymbol{m} \in \{0,1\}^L$ zeroes out padded positions by adding a large negative constant $-C$ ($C \gg 0$):

$$\tilde{e}_i = e_i + (1 - m_i)(-C) \tag{14}$$

followed by a softmax over i:

$$\alpha_i = \frac{\exp(\tilde{e}_i)}{\sum_{j=1}^{L}\exp(\tilde{e}_j)}, \qquad \sum_{i=1}^{L}\alpha_i = 1 \tag{15}$$

The pooled variant representation is:

$$\boldsymbol{z} = \sum_{i=1}^{L} \alpha_i \boldsymbol{h}_i \in \mathbb{R}^d \tag{16}$$

This pooling is applied independently to wild-type and mutant features:

$$\boldsymbol{z}_{\mathrm{WT}} = \mathrm{Pool}(\boldsymbol{H})_{\mathrm{WT}},\ \boldsymbol{z}_{\mathrm{Mut}} = \mathrm{Pool}(\boldsymbol{H})_{\mathrm{Mut}}$$

which are subsequently fed into the prediction head (see next subsection). In ablation experiments, replacing eqn (16) with mean or max pooling revealed trade-offs between correlation and error, where attention pooling yielded more balanced performance across mutation types.

4.7 Prediction of the stability change

Given pooled representations of wild type and mutant, $\boldsymbol{z}_{\mathrm{WT}}, \boldsymbol{z}_{\mathrm{Mut}} \in \mathbb{R}^{d_{\mathrm{feat}}}$, we form a translation-invariant difference to isolate mutation-induced changes:

$$\Delta\boldsymbol{z} = \boldsymbol{z}_{\mathrm{Mut}} - \boldsymbol{z}_{\mathrm{WT}} \tag{17}$$

A regression head $h_\phi : \mathbb{R}^{d_{\mathrm{feat}}} \to \mathbb{R}$ then predicts the stability change:

$$\widehat{\Delta\Delta G} = h_\phi(\Delta\boldsymbol{z}) \tag{18}$$

We instantiate h_ϕ as a lightweight MLP with ReLU activations, batch normalization, and dropout between layers. Full architectural hyperparameters are provided in the SI.

4.8 Model training

We optimize the regression objective on experimental $\Delta\Delta G$ with mean absolute error (MAE):

$$\mathcal{L}_{\mathrm{MAE}} = \frac{1}{N}\sum_{n=1}^{N}\left|\widehat{\Delta\Delta G}_n - \Delta\Delta G_n\right| \tag{19}$$

Models are trained with AdamW (learning rate 1×10^{-4}), a ReduceLROnPlateau scheduler, and early stopping on validation MAE (up to 50 epochs; patience = 10). Training uses distributed data parallelism on NVIDIA GeForce RTX 4090 GPUs with fixed random seeds. Full hyperparameters (batch size, gradient accumulation, weight decay, precision, and data loader settings) are provided in the SI.

4.9 Feature and geometry computations

We computed a set of structural and geometric descriptors to interpret model predictions and assess performance across mutation contexts.

4.9.1 Solvent-accessible surface area (SASA). Residue-level SASA values were calculated using the Shrake–Rupley algorithm (Bio.PDB) on wild-type structures, with a probe radius of 1.4 Å.[53] Mutations were classified as *core* (SASA < 25 Å^2)[54] or *surface* (SASA ≥ 25 Å^2), and double mutants grouped into core–core, core–surface, or surface–surface categories. Performance metrics (PCC, SCC, RMSE, MAE) were computed separately for each category.

4.9.2 Inter-mutation Cα distance. For double mutants, the Euclidean distance between Cα atoms of the mutated residues was computed from wild-type structures, and categorized as *close* (≤5 Å), *medium* (5–10 Å), or *distant* (>10 Å).[55,56] Category-specific performance was compared, and statistical significance was assessed *via* two-sided Mann–Whitney *U* tests.

4.9.3 Feature-space distance. For each wild-type–mutant pair, pooled representations were standardized per feature (z-score) and the feature-space distance was defined as

$$d_{\text{feat}} = \|\hat{\boldsymbol{z}}_{\text{Mut}} - \hat{\boldsymbol{z}}_{\text{WT}}\|_2 \tag{20}$$

where $\hat{\boldsymbol{z}} = (\boldsymbol{z} - \boldsymbol{\mu})/\boldsymbol{\sigma}$ denotes per-feature normalization. Distances were analyzed as a function of mutation count and protein identity to test the hypothesis that more mutations induce larger representational shifts.

4.9.4 Local RMSD. Mutant structures were predicted by Chai-1 (ref. 45) and compared to wild-type structures. Local RMSD was calculated within 4 Å and 8 Å radii around mutation sites using the Bio.PDB Superimposer. The correlation between local RMSD and feature-space distance was used to assess structural interpretability.

Unless otherwise stated, residue–residue interactions in attention analyses were defined by a minimum heavy-atom distance ≤3.5 Å.[57]

Author contributions

Conceptualization: H. T., S. L.; methodology: H. T., S. L.; formal analysis: H. T.; writing – original draft: H. T.; writing – review & editing: H. T., S. L., Y. X.; supervision: Y. X.; funding acquisition: Y. X.

Conflicts of interest

The authors declare that they have no competing interests.

Data availability

The source code for the UniStab model is available at https://github.com/xlab-BioAI/UniStab. The datasets used for training and evaluation can be accessed at https://github.com/xlab-BioAI/UniStab/tree/main/data.

Supplementary information (SI): methodological details, ROC and precision-recall analyses, and supporting results. See DOI: https://doi.org/10.1039/d6sc02103d.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (Grant No. 2023YFC2506402 and 2024YFA1306902), the National Natural Science Foundation of China (Grant No. 32570779), and the Shanghai Municipal Education Commission (Grant No. 2024AIZD008). The authors thank Henry Dieckhaus for assistance with ThermoMPNN-D benchmarking and valuable discussions on model evaluation.

Notes and references

† https://protddg-bench.github.io/.

1 N. Chennamsetty, V. Voynov, V. Kayser, B. Helk and B. L. Trout, Design of therapeutic proteins with enhanced stability, *Proc. Natl. Acad. Sci.*, 2009, **106**, 11937–11942.
2 A. Tennenhouse, *et al.*, Computational optimization of antibody humanness and stability by systematic energy-based ranking, *Nat. Biomed. Eng.*, 2024, **8**, 30–44.
3 B. K. Shoichet, W. A. Baase, R. Kuroki and B. W. Matthews, A relationship between protein stability and protein function, *Proc. Natl. Acad. Sci.*, 1995, **92**, 452–456.
4 R. Vanella, *et al.*, Understanding activity-stability tradeoffs in biocatalysts by enzyme proximity sequencing, *Nat. Commun.*, 2024, **15**, 1807.
5 F. Jiang, *et al.*, A general temperature-guided language model to design proteins of enhanced stability and activity, *Sci. Adv.*, 2024, **10**, eadr2641.
6 J. Zheng, N. Guo and A. Wagner, Selection enhances protein evolvability by increasing mutational robustness and foldability, *Science*, 2020, **370**, eabb5962.
7 L. M. Blaabjerg, *et al.*, Rapid protein stability prediction using deep learning representations, *Elife*, 2023, **12**, e82593.
8 Y. Zhou, Q. Pan, D. E. Pires, C. H. Rodrigues and D. B. Ascher, DDMut: predicting effects of mutations on protein stability using deep learning, *Nucleic Acids Res.*, 2023, **51**, W122–W128.
9 J. B. Kinney and D. M. McCandlish, Massively parallel assays and quantitative sequence–function relationships, *Annu. Rev. Genom. Hum. Genet.*, 2019, **20**, 99–127.
10 D. M. Fowler and S. Fields, Deep mutational scanning: a new style of protein science, *Nat. Methods*, 2014, **11**, 801–807.
11 D. T. Dryden, A. R. Thomson and J. H. White, How much of protein sequence space has been explored by life on Earth?, *J. R. Soc., Interface*, 2008, **5**, 953–956.
12 C. A. Olson, N. C. Wu and R. Sun, A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain, *Curr. Biol.*, 2014, **24**, 2643–2651.
13 Y. Liu, *et al.*, Enhancing functional proteins through multimodal inverse folding with ABACUS-T, *Nat. Commun.*, 2025, **16**, 10177.

14 J. Schymkowitz, *et al.*, The FoldX web server: an online force field, *Nucleic Acids Res.*, 2005, **33**, W382–W388.
15 E. H. Kellogg, A. Leaver-Fay and D. Baker, Role of conformational sampling in computing mutation-induced changes in protein structure and stability, *Proteins: Struct., Funct., Bioinf.*, 2011, **79**, 830–838.
16 D. Umerenkov, *et al.*, PROSTATA: a framework for protein stability assessment using transformers, *Bioinformatics*, 2023, **39**, btad671.
17 C. Savojardo, M. Manfredi, P. L. Martelli and R. Casadio, DDGemb: predicting protein stability change upon single- and multi-point variations with embeddings and deep learning, *Bioinformatics*, 2025, **41**, btaf019.
18 Z. Nie, *et al.*, Predicting protein stability changes upon mutations with dual-view ensemble learning from single sequence, *Briefings Bioinf.*, 2025, **26**, bbaf319.
19 Z. Lin, *et al.*, Evolutionary-scale prediction of atomic-level protein structure with a language model, *Science*, 2023, **379**, 1123–1130.
20 D. J. Diaz, *et al.*, Stability Oracle: a structure-based graph-transformer framework for identifying stabilizing mutations, *Nat. Commun.*, 2024, **15**, 6170.
21 J. Sun, T. Zhu, Y. Cui and B. Wu, Structure-based self-supervised learning enables ultrafast protein stability prediction upon mutation, *Innovation*, 2025, **6**, 100750.
22 Y. Chen, Y. Xu, D. Liu, Y. Xing and H. Gong, An end-to-end framework for the prediction of protein structure and fitness from single sequence, *Nat. Commun.*, 2024, **15**, 7400.
23 J. Ouyang-Zhang, D. J. Diaz, A. R. Klivans and P. Krähenbühl, Predicting a protein's stability under a million mutations, *Adv. Neural Inf. Process. Syst.*, 2023, **36**, 76229–76247.
24 J. Dauparas, *et al.*, Robust deep learning–based protein sequence design using ProteinMPNN, *Science*, 2022, **378**, 49–56.
25 Z. Gao, C. Tan, P. Chacón and S. Z. Li, PiFold: Toward effective and efficient protein inverse folding, *arXiv*, 2022, preprint, arXiv:2209.12643, DOI: 10.48550/arXiv.2209.12643.
26 P. Bai, *et al.*, Mask-prior-guided denoising diffusion improves inverse protein folding, *Nat. Mach. Intell.*, 2025, **7**, 876–888.
27 H. Dieckhaus, M. Brocidiacono, N. Z. Randolph and B. Kuhlman, Transfer learning to leverage larger datasets for improved prediction of protein stability changes, *Proc. Natl. Acad. Sci.*, 2024, **121**, e2314853121.
28 Z. Li and Y. Luo, Generalizable and scalable protein stability prediction with rewired protein generative models, *Nat. Commun.*, 2026, **17**, 891.
29 H. Tan *et al.*, ProStab: Prediction of protein stability change upon mutations by protein language and inverse folding models, 2025, *bioRxiv*, preprint, 2025.08.11.669595, DOI: 10.1101/2025.08.11.669595.
30 H. Dieckhaus and B. Kuhlman, Protein stability models fail to capture epistatic interactions of double point mutations, *Protein Sci.*, 2025, **34**, e70003.
31 M. Cagiada, S. Ovchinnikov and K. Lindorff-Larsen, Predicting absolute protein folding stability using generative models, *Protein Sci.*, 2025, **34**, e5233.
32 J. Jumper, *et al.*, Highly accurate protein structure prediction with AlphaFold, *Nature*, 2021, **596**, 583–589.
33 J. Abramson, *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3, *Nature*, 2024, **630**, 493–500.
34 B. Li, Y. T. Yang, J. A. Capra and M. B. Gerstein, Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks, *PLoS Comput. Biol.*, 2020, **16**, e1008291.
35 Y. Xu, D. Liu and H. Gong, Improving the prediction of protein stability changes upon mutations by geometric learning and a pre-training strategy, *Nat. Comput. Sci.*, 2024, **4**, 840–850.
36 S. Wang, H. Tang, Y. Zhao and L. Zuo, BayeStab: Predicting effects of mutations on protein stability with uncertainty quantification, *Protein Sci.*, 2022, **31**, e4467.
37 F. Zheng, Y. Liu, Y. Yang, Y. Wen and M. Li, Assessing computational tools for predicting protein stability changes upon missense mutations using a new dataset, *Protein Sci.*, 2024, **33**, e4861.
38 S. Reeves and S. Kalyaanamoorthy, Zero-shot transfer of protein sequence likelihood models to thermostability prediction, *Nat. Mach. Intell.*, 2024, **6**, 1063–1076.
39 J. Xu *et al.*, InstructPLM-mu: 1-hour fine-tuning of ESM2 beats ESM3 in protein mutation predictions, 2025, *arXiv*, preprint, arXiv:2510.03370, DOI: 10.48550/arXiv.2510.03370.
40 M. Topolska, A. Beltran and B. Lehner, Deep indel mutagenesis reveals the impact of amino acid insertions and deletions on protein stability and function, *Nat. Commun.*, 2025, **16**, 2617.
41 S. Savino, T. Desmet and J. Franceus, Insertions and deletions in protein evolution and engineering, *Biotechnol. Adv.*, 2022, **60**, 108010.
42 J. E. Hu *et al.*, LoRA: Low-rank adaptation of large language models, 2021, *arXiv*, preprint, arXiv:2106.09685, DOI: 10.48550/arXiv.2106.09685.
43 K. Tsuboyama, *et al.*, Mega-scale experimental analysis of protein folding stability in biology and design, *Nature*, 2023, **620**, 434–444.
44 L. Montanucci, *et al.*, DDGun: an untrained predictor of protein stability changes upon amino acid variants, *Nucleic Acids Res.*, 2022, **50**, W222–W227.
45 C. Discovery *et al.*, Chai-1: Decoding the molecular interactions of life, *bioRxiv*, 2024, 2024.10.10.615955, DOI: 10.1101/2024.10.10.615955.
46 A. J. Faure, *et al.*, The genetic architecture of protein stability, *Nature*, 2024, **634**, 995–1003.
47 T. Mi *et al.*, Accelerating virtual directed evolution of proteins via reinforcement learning, *bioRxiv*, 2025, preprint, 2025.06.25.661516, DOI: 10.1101/2025.06.25.661516.
48 H. Sun, *et al.*, Accelerating protein engineering with fitness landscape modelling and reinforcement learning, *Nat. Mach. Intell.*, 2025, **7**, 1446–1460.
49 S. Alamdari *et al.*, Protein generation with evolutionary diffusion: sequence is all you need, 2024, *bioRxiv*, preprint, 2023.09.11.556673, DOI: 10.1101/2023.09.11.556673.

50 P. Notin, N. Rollins, Y. Gal, C. Sander and D. Marks, Machine learning for functional protein design, *Nat. Biotechnol.*, 2024, **42**, 216–228.
51 T. Hayes, *et al.*, Simulating 500 million years of evolution with a language model, *Science*, 2025, **387**, 850–858.
52 L. Montanucci, E. Capriotti, Y. Frank, N. Ben-Tal and P. Fariselli, DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations, *BMC Bioinf.*, 2019, **20**, 335.
53 A. Shrake and J. A. Rupley, Environment and exposure to solvent of protein atoms. Lysozyme and insulin, *J. Mol. Biol.*, 1973, **79**, 351–371.
54 S. Iqbal, *et al.*, MISCAST: MIssense variant to protein StruCture Analysis web SuiTe, *Nucleic Acids Res.*, 2020, **48**, W132–W139.
55 I. Anishchenko, S. Ovchinnikov, H. Kamisetty and D. Baker, Origins of coevolution between residues distant in protein 3D structures, *Proc. Natl. Acad. Sci.*, 2017, **114**, 9122–9127.
56 C. Yuan, H. Chen and D. Kihara, Effective inter-residue contact definitions for accurate protein fold recognition, *BMC Bioinf.*, 2012, **13**, 292.
57 A. L. Cuff, R. W. Janes and A. C. Martin, Analysing the ability to retain sidechain hydrogen-bonds in mutant proteins, *Bioinformatics*, 2006, **22**, 1464–1470.