

Multi-task deep kernel learning for uncertainty-aware prediction of protein stability changes

Jiameng Hu^{1,2}, Chao Wang^{2#}

¹ School of Life Science and Technology, HIT Faculty of Life Science and Medicine, Harbin Institute
of Technology, Harbin, China

² Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen, China

[#] Correspondence: chaowang@szbl.ac.cn

Abstract

Accurate prediction of mutation-induced stability changes is crucial for interpreting variant effects and guiding protein engineering, but integrating heterogeneous measurements and quantifying uncertainty remain challenging. Here we present STAB-DKL, an uncertainty-aware multi-task deep kernel learning framework that combines curated and megascale folding free-energy changes ($\Delta\Delta G$) with melting-temperature changes (ΔT_m) as auxiliary supervision. Selective representation sharing and dataset-specific Gaussian process (GP) heads enable joint learning across datasets. The framework integrates pretrained sequence and structural representations with StrucMSA, a compact and computationally efficient evolutionary representation incorporating local structural context. Across four held-out benchmarks, STAB-DKL achieved higher average predictive correlations than the evaluated methods and provided informative uncertainty estimates. On a retrospective seven-protein benchmark, it improved ranking of stabilizing mutations (NDCG@5) by 68% over ThermoMPNN, the strongest comparator. Together, these results demonstrate how multi-task fusion and dataset-specific probabilistic modeling of heterogeneous data enable accurate, uncertainty-aware protein stability prediction and stabilizing mutation prioritization.

Main

Protein function depends on the ability of polypeptide chains to fold into stable native conformations while avoiding misfolded or aggregated states¹. At the same time, evolution selects protein sequences for function rather than simply for maximal stability, and many natural proteins are only marginally stable²⁻⁴. A single amino acid substitution can therefore destabilize the native state and perturb the folding equilibrium, increasing the risk of misfolding and aggregation. Such stability changes contribute to the molecular basis of many disease-associated variants⁵⁻⁸ and pose major challenges for protein engineering⁹⁻¹². Accurate prediction of mutation-induced stability changes is thus important for interpreting variant effects in human disease and for prioritizing stabilizing mutations in enzyme engineering and protein design.

Mutation-induced stability changes are commonly quantified by $\Delta\Delta G$, which represents the difference in folding free energy between a mutant protein and its wild-type counterpart. Experimental measurement of $\Delta\Delta G$ and related thermodynamic parameters remains labor-intensive and difficult to scale across diverse proteins and mutational landscapes, highlighting the value of computational approaches for stability prediction¹³. Physics-based methods, including empirical energy functions and molecular modeling approaches, provide useful mechanistic insights^{14,15}. However, their quantitative accuracy remains sensitive to energy-function approximations, conformational sampling, and input structural models. Their scalability to high-throughput mutational screening is further constrained by computational cost. These limitations have motivated the development of machine learning models trained on experimental stability measurements.

Early machine learning models for protein stability prediction relied mainly on manually curated thermodynamic datasets, which are relatively small but cover diverse protein lengths, folds and experimental conditions¹⁶. More recently, megascale stability datasets have greatly expanded the number of available measurements, but these data are enriched for short sequences and single-domain proteins and cover a relatively restricted stability range¹⁷. A growing number of protein stability predictors have demonstrated the value of megascale measurements¹⁸⁻²². However, limited overall gains from incorporating conventional thermodynamic measurements have been reported^{18,20}. Alongside these $\Delta\Delta G$ measurements, changes in melting temperature (ΔT_m) provide a related but non-

equivalent measure of protein stability²¹. Effectively integrating these data sources, while accounting for their differences in scale, coverage and experimental noise, remains a central challenge for improving mutation-induced stability prediction.

Effectively learning from these heterogeneous datasets also depends on how protein variants are represented. Data-driven approaches initially relied on physicochemical descriptors and MSA-derived evolutionary features²³⁻²⁶, whereas recent models increasingly exploit pretrained sequence and structure representations^{18-20,22,27,28}. These feature types capture complementary information about sequence context, structural features and evolutionary constraints, but how best to integrate them remains incompletely understood. Particularly, conventional MSA-derived features primarily reflect sequence conservation without explicitly accounting for the three-dimensional environment surrounding the mutated residue. Incorporating local structural context into evolutionary representations may therefore provide more relevant information for predicting mutation-induced stability changes.

Reliable protein stability prediction requires both accurate estimation of mutation effects and quantification of predictive uncertainty, particularly when experimental measurements are limited, heterogeneous and noisy. Many widely used predictors provide point estimates without indicating how uncertainty varies across mutations. Gaussian processes (GPs) offer a probabilistic framework in which covariance functions encode similarity between observations and predictions are expressed as posterior predictive distributions²⁹⁻³⁴. However, conventional kernels operating on fixed protein representations may be limited in capturing the relationships most relevant to protein stability, while exact GP inference scales poorly with dataset size. Deep kernel learning (DKL) increases model flexibility by jointly learning a nonlinear feature transformation and GP kernel parameters³⁵, while structured kernel interpolation enables scalable approximate inference³⁶. Within a multi-task framework³⁷, shared representations enable information transfer, while separate GP heads and likelihoods allow dataset-specific covariance and noise parameters. This combination provides a probabilistic basis for integrating complementary stability measurements while accommodating differences between datasets.

For protein engineering, evaluation should also reflect the limited number of variants that can be tested for each target³⁸. Correlations and error metrics summarize agreement across measured

mutations but do not directly quantify the recovery of stabilizing variants among a small set of top-ranked candidates. Protein-level evaluation should therefore consider both the proportion of selected mutations that are stabilizing and whether those with larger stabilizing effects are ranked earlier. Such analyses can reveal differences in practical utility that are not fully captured by overall predictive performance.

In this work, we introduce STAB-DKL, a multi-task deep kernel learning framework for protein stability prediction and stabilizing mutation prioritization. STAB-DKL jointly learns from curated and megascale $\Delta\Delta G$ measurements through shared representations and dataset-specific GP heads, while incorporating ΔT_m as auxiliary supervision through selective representation sharing. The framework integrates sequence, structural and evolutionary representations, including a structure-constrained MSA descriptor designed to capture evolutionary constraints within the local structural environment of the mutated residue. We evaluate predictive accuracy and uncertainty across held-out benchmarks and assess mutation prioritization using a retrospective protein-level ranking benchmark. These analyses show that integrating complementary stability measurements and protein representations through STAB-DKL improves average predictive performance and the recovery of stabilizing mutations while providing informative uncertainty estimates.

Results

STAB-DKL integrates heterogeneous stability information within a multi-task deep kernel learning framework

To integrate heterogeneous protein stability data and quantify predictive uncertainty, we developed STAB-DKL, a multi-task deep kernel learning framework for mutation-induced stability prediction (**Fig. 1**). The framework was designed around three key aspects: data integration, feature representation, and model architecture.

At the data level (**Fig. 1**, input data), megascale and curated $\Delta\Delta G$ datasets differ substantially in sampling density, protein coverage and experimental characteristics. ThermoMPNN¹⁸ and Stability Oracle²⁰ previously evaluated training strategies that combine megascale and curated $\Delta\Delta G$ measurements but reported limited overall benefit under their respective training protocols. To address this, STAB-DKL models the two data sources within a multi-task framework that enables shared

learning while retaining dataset-specific modeling. We further incorporate ΔT_m as an auxiliary task to provide complementary, but non-equivalent, stability supervision.

At the feature level (**Fig. 1**, feature extractor), STAB-DKL integrates complementary pretrained sequence, structure-informed and alignment-based evolutionary representations. ESM2³⁹ embeddings provide contextual sequence representations that implicitly encode evolutionary constraints learned from large-scale natural protein sequences. For structural representation, we extended the ThermoMPNN¹⁸ feature-extraction strategy by deriving ProteinMPNN⁴⁰ representations under both backbone-only and backbone-plus-sequence conditioning, allowing geometry-dominated and sequence-conditioned structural information to be captured separately. For evolutionary information, we developed a structure-constrained MSA representation (StrucMSA) that incorporates the local three-dimensional environment of the mutated residue into alignment-based evolutionary features (see below and methods).

At the architectural level (**Fig. 1**, model architecture), STAB-DKL combines multi-task representation learning with deep kernel learning. We adapted the Light Attention (LA) module⁴¹ used in ThermoMPNN¹⁸ to transform and integrate heterogeneous input features. The curated and megascale $\Delta\Delta G$ branches share this LA module and a multilayer perceptron (MLP) module to learn common stability-related representations, while retaining separate Gaussian process (GP) heads for dataset-specific probabilistic prediction. The ΔT_m auxiliary task shares only the LA module but uses its own MLP and GP head. During joint training, supervision from this dedicated ΔT_m branch also updates the shared LA parameters. Each GP head produces both a predictive mean and variance, with Kernel Interpolation for Scalable Structured Gaussian Processes (KISS-GP)³⁶ enabling scalable inference. This architecture supports joint learning across heterogeneous stability datasets while allowing each GP head to learn its own covariance and observation-noise parameters.

Heterogeneous stability data provide complementary information

To support multi-task learning across heterogeneous protein stability data, we assembled candidate data from curated $\Delta\Delta G$, megascale $\Delta\Delta G$, and ΔT_m measurements (**Fig. 2a**). Before constructing the final training sets, we predefined four held-out benchmarks for model evaluation (**Fig. 2a**; Test set): the multi-protein Ssym⁴² and S669¹⁶ datasets and the single-protein myoglobin⁴³ and p53⁴⁴ datasets. S669 was derived from ThermoMutDB⁴⁵ and curated to reduce redundancy with

commonly used training sets, whereas Ssym is a ProTherm-derived⁴⁶ benchmark containing matched direct and reverse mutations. Because some benchmark records share source databases with the training pool, and because sequence homology could introduce information leakage, we fixed the benchmark sets in advance and excluded any candidate training protein sharing more than 30% sequence identity with a benchmark protein.

Following this filtering, the three training sources were curated separately (**Fig. 2a**). For the curated $\Delta\Delta G$ dataset, we integrated the widely used S2648⁴⁷, Q3488⁴⁸ and S9028⁴⁹ datasets with records from ThermoMutDB⁴⁵ and FireProtDB⁵⁰. After reconciling overlapping entries and applying the sequence filter, 4,238 mutations were retained, forming the C4238 dataset. Experimental ΔT_m records from ThermoMutDB⁴⁵ and FireProtDB⁵⁰ were processed using the same sequence-leakage criterion, yielding 2,403 mutations (Tm2403). To minimize sequence similarity between the training and validation sets, the remaining megascale proteins were clustered at 30% sequence identity, and 20 proteins from singleton clusters were reserved as the Ms20 validation set. The remaining 154 proteins formed the Ms154 training set.

We next compared the three training datasets in terms of protein coverage, mutational density and stability distributions to assess their complementarity. Ms154 provides substantially denser mutational coverage per protein than C4238 and Tm2403 (**Fig. 2b**) but is concentrated in shorter sequences (**Fig. 2c**), with a median length of 56 residues compared with 131 residues for C4238 and 199 residues for Tm2403 (**Fig. 2c**). The megascale $\Delta\Delta G$ measurements also span a narrower range than the curated $\Delta\Delta G$ data (**Supplementary Fig. 1a**). Sequence-space analysis showed that C4238 and Ms154 have limited similarity, with only 8% of C4238 sequences sharing more than 30% sequence identity with Ms154 sequences (**Fig. 2d**). Tm2403 further broadens the sequence diversity represented in the training data (**Fig. 2d**). This complementarity was also evident in mutation coverage, with 71% of C4238 mutations and 94% of Tm2403 mutations occurring in proteins without a sequence match in Ms154 (**Supplementary Fig. 1b**). In addition, ΔT_m and $\Delta\Delta G$ showed a strong inverse correlation ($r = -0.76$; **Supplementary Fig. 1c**), consistent with ΔT_m providing related but non-equivalent information about protein stability. Together, these differences support the complementary use of curated $\Delta\Delta G$, megascale $\Delta\Delta G$ and ΔT_m in the multi-task framework.

Because Ms154 was the largest $\Delta\Delta G$ training source and was strongly enriched for destabilizing

mutations (**Supplementary Fig. 1a**), we evaluated thermodynamic reversibility (TR) and thermodynamic permutation (TP) to improve representation of stabilizing effects (**Supplementary Fig. 2a**). Without augmentation, mean Pearson correlations were 0.708 for direct mutations and 0.478 for reverse mutations (**Supplementary Fig. 2b**). Applying TR to both the curated and megascale datasets increased the reverse-mutation correlation to 0.707 while maintaining comparable performance for direct mutations ($r = 0.706$; **Supplementary Fig. 2b**). This is consistent with previous findings that reverse-mutation augmentation can reduce prediction bias and improve the prediction of stabilizing variants^{28,48,51}. For the curated datasets, TR generally outperformed TP+TR, whereas adding TP to TR for Ms154 yielded small improvements in Pearson correlation with similar Spearman correlations. We therefore used TR for C4238 and Tm2403 and TP+TR for Ms154. Although TP+TR expanded Ms154 to approximately 2.84 million examples (**Supplementary Fig. 2c**), predictive correlations approached a plateau at around 20,000 sampled examples, with only small additional gains at larger sample sizes (**Supplementary Fig. 2d**).

We next assessed the contribution of each supervision source by systematically removing individual datasets from the complete model. Excluding C4238, Ms154 or Tm2403 reduced the mean Pearson correlation by approximately 0.03 in each case (**Fig. 2e**), demonstrating that curated $\Delta\Delta G$, megascale $\Delta\Delta G$ and ΔT_m each provide measurable and complementary information for stability prediction. Whereas previous studies reported limited gains from direct pooling or sequential training^{18,20}, STAB-DKL improved stability prediction by integrating heterogeneous data sources within a multi-task framework.

Multimodal feature integration improves protein stability prediction

After establishing the complementary contributions of the three training data sources, we next turned to feature representation and examined how sequence, structural and evolutionary information contribute to stability prediction. The complete representation comprised 1,280-dimensional ESM2 features, 1,066-dimensional ProteinMPNN features and 61-dimensional MSA-derived features (**Supplementary Fig. 3a-b**). ProteinMPNN features were organized according to two complementary groupings: decoding condition, comparing backbone-only (MPNN-BB) with backbone-plus-sequence conditioning (MPNN-BB&SQ), and feature type, comparing learned embeddings (MPNN-Emb) with amino-acid log-probability features (MPNN-LogP) (**Supplementary Fig. 3a**). For MSA-derived

representations, we first included conventional sequence-based features (MSA-Seq). We further developed StrucMSA, which incorporates the local three-dimensional environment of the mutated residue into alignment-derived evolutionary information (**Supplementary Fig. 3c**; see Methods). The resulting 60-dimensional feature set is referred to as MSA-Struc in subsequent analyses (**Supplementary Fig. 3b**).

We first characterized the associations between individual feature groups and experimentally measured stability effects across the four test datasets (**Fig. 3a**). These associations varied across benchmarks. ProteinMPNN-derived features generally showed the strongest individual correlations, whereas ESM-derived representations were particularly informative for myoglobin (**Fig. 3a**). When the ProteinMPNN features were grouped by decoding condition, MPNN-BB&SQ generally showed stronger associations than MPNN-BB (**Fig. 3a**). When grouped by feature type, MPNN-LogP generally showed stronger correlations than MPNN-Emb (**Fig. 3a**). These patterns indicate that both sequence conditioning and the type of ProteinMPNN-derived representation influence their association with mutation-induced stability effects.

We then quantified the contribution of these representations through feature ablation (**Fig. 3b**). Removing the complete ProteinMPNN feature set produced the largest performance decrease, reducing the mean Pearson correlation from 0.713 to 0.618 for direct mutations and from 0.712 to 0.617 for reverse mutations (**Fig. 3b**, upper panel). Removing ESM features also substantially reduced performance, with Pearson correlation decreasing to 0.646 and 0.643 for direct and reverse mutations respectively (**Fig. 3b**, upper panel). Similar trends were observed for Spearman correlation (**Fig. 3b**, lower panel), supporting complementary contributions from pretrained sequence and structure-informed representations. We next dissected the ProteinMPNN contribution according to the same two feature groupings (**Fig. 3b**). Removing MPNN-Emb caused the largest decrease among the ProteinMPNN subsets, lowering the mean Pearson correlation by approximately 0.035, whereas removing MPNN-LogP resulted in a smaller decrease of approximately 0.01–0.02. Similarly, ablating either the backbone-only or backbone-plus-sequence-conditioned features produced modest reductions in performance. The magnitude of these effects varied across benchmarks, indicating that different components of the ProteinMPNN representation contribute differently across protein contexts.

We next examined the MSA-derived representations in greater detail. Although removal of the

complete MSA feature set reduced model performance, this effect was driven primarily by MSA-Struc (**Fig. 3b**). Removing MSA-Struc decreased the mean Pearson correlation from 0.713 to 0.688 for direct mutations and from 0.712 to 0.686 for reverse mutations, whereas removal of the conventional sequence-based MSA features (MSA-Seq) produced little change (**Fig. 3b**). Consistently, MSA-Struc showed stronger associations with experimental stability effects than MSA-Seq in three of the four benchmark datasets, with the largest difference observed for myoglobin (**Fig. 3a**). Pairwise feature correlation analysis further revealed substantial shared information between MSA-Struc and MSA-Seq (**Supplementary Fig. 3d**). Together, these results suggest that incorporating local structural context enhances the predictive utility of alignment-derived evolutionary information while retaining substantial information shared with conventional MSA representations.

To compare the predictive utility of MSA-Struc with a pretrained MSA representation, we replaced MSA features with embeddings from the MSA Transformer (ESM-MSA-1b). The two representations yielded comparable Pearson correlations, whereas the MSA Transformer produced a modestly lower mean Spearman correlation by approximately 0.015 (**Fig. 3c**). Combining both representations did not consistently improve prediction (**Fig. 3d**) and canonical correlation analysis yielded correlations of 0.89 and 0.78 for the first two pairs of canonical components (**Supplementary Fig. 3e**). These results suggest that MSA-Struc provides predictive utility comparable to that of pretrained MSA embeddings within STAB-DKL while requiring no neural-network pretraining or GPU inference. It also has substantially lower dimensionality and was generated 18 to 42 times faster in our implementation (**Fig. 3e**). Together, these results establish structure-constrained MSA as an efficient means of incorporating structurally informed evolutionary information into protein stability prediction.

Multi-task deep kernel learning enables dataset-aware probabilistic modeling

We next examined the architectural choices underlying STAB-DKL and how information should be shared across heterogeneous stability tasks. Removing the shared LA-derived module reduced Pearson correlation by 0.072 and 0.025 and Spearman correlation by 0.058 and 0.023 for direct and reverse mutations, respectively (**Fig. 4a**). These results support the contribution of this learned feature transformation to stability prediction from heterogeneous sequence, structural, and evolutionary representations.

We then evaluated how MLP sharing affects predictive performance (**Fig. 4a**). Assigning separate MLPs to the C4238 and Ms154 $\Delta\Delta G$ branches reduced Pearson correlation by approximately 0.03 in both mutation directions, supporting the benefit of learning a common representation across these datasets. Extending MLP sharing to the auxiliary ΔT_m branch produced a larger reduction of approximately 0.04~0.05, suggesting that using a separate MLP for the auxiliary ΔT_m task is more effective than sharing a single MLP across all three branches. Spearman correlation showed similar trends across these ablations. Together, these results support sharing an MLP between the two $\Delta\Delta G$ branches while retaining a separate MLP for ΔT_m .

STAB-DKL combines neural representation learning with GP regression within a DKL framework to support probabilistic stability prediction and uncertainty estimation. We first assessed the predictive contribution of this DKL formulation. Removing the GP layer while retaining the upstream neural architecture reduced Pearson correlation by approximately 0.025 (**Fig. 4a**), supporting the benefit of combining learned representations with GP regression. We then examined whether the two $\Delta\Delta G$ datasets benefit from separate GP heads. Although the curated and megascale branches share the upstream LA-derived module and MLP, the joint-task variant reduced mean Pearson correlation from 0.713 to 0.694 for direct mutations and from 0.712 to 0.690 for reverse mutations, with smaller reductions in mean Spearman correlation in both directions (**Fig. 4a**). These results favor sharing representations across the two $\Delta\Delta G$ datasets while retaining dataset-specific GP heads. The predictive distributions from these heads are subsequently combined through precision-weighted fusion (**Fig. 1**), with each head's contribution determined by the inverse of its predictive variance.

Having established the predictive benefit of the DKL architecture, we next evaluated its uncertainty estimates against two alternatives: Monte Carlo (MC) dropout⁵² and a probabilistic regression model trained using Gaussian negative log-likelihood (Gaussian NLL)^{53,54}. STAB-DKL achieved higher overall predictive correlations, exceeding MC dropout and Gaussian NLL by approximately 0.02 and 0.06 in Pearson correlation, respectively (**Fig. 4b**). To assess uncertainty quality, we used miscalibration area, NLL, and the correlation between prediction residuals and predicted variance. STAB-DKL achieved lower miscalibration area and NLL than MC dropout across all evaluated datasets, whereas Gaussian NLL regression yielded the lowest values for both metrics (**Supplementary Fig. 4a**). Residual–variance correlations varied across datasets, with no method

consistently performing best (**Supplementary Fig. 4a**). Together, these results support our choice of GP-based DKL for its higher predictive accuracy while retaining informative uncertainty estimates.

We then examined whether predicted variance could identify more reliable predictions. Selecting progressively lower-variance subsets reduced RMSE and MAE across all test datasets (**Fig. 4c**). For p53, retaining the lowest-variance 20 to 40% of predictions reduced MAE to approximately 40% of that for the full test set (**Fig. 4c**). These results suggest that predicted variance provides a useful criterion for selecting more reliable stability predictions. Low-variance predictions clustered around predicted $\Delta\Delta G$ values near zero, whereas larger-magnitude predictions generally exhibited higher variance (**Supplementary Fig. 4b**). Given the greater representation of near-zero $\Delta\Delta G$ values in the training data, this pattern suggests that the reliability of selected predictions may partly reflect training-data coverage across the stability range.

Together, these analyses support sharing representations between the two $\Delta\Delta G$ branches while retaining dataset-specific GP heads and a separate MLP for the auxiliary ΔT_m task. This DKL architecture combines improved predictive performance with uncertainty estimates that can guide prediction prioritization.

STAB-DKL achieves strong overall performance across diverse benchmarks

We evaluated STAB-DKL on four benchmark datasets spanning multi-protein collections (Ssym and S669) and individual proteins (myoglobin and p53). Model selection was based on validation-set performance, and STAB-DKL results were averaged over three random seeds. PROSTATA²⁸ and ThermoMPNN¹⁸ were retrained using our training data, while published results were used for Stability Oracle²⁰ because its training code was unavailable (**Fig. 5**).

STAB-DKL achieved the highest mean Pearson and Spearman correlations in both mutation directions, reaching approximately 0.71 for both metrics and exceeding Stability Oracle by approximately 0.05 and 0.07, respectively (**Fig. 5**). The largest Pearson gains were observed on p53, where correlations reached 0.85 and 0.86 for direct and reverse mutations, exceeding the strongest comparator in each direction by 0.09 and 0.13, respectively. On myoglobin, Spearman correlation reached 0.73 in both directions, compared with 0.66 for Stability Oracle. On S669, STAB-DKL achieved the highest Pearson correlation of 0.56 and matched the retrained ThermoMPNN in Spearman

correlation (0.58) in both directions. On Ssym, Pearson correlations were slightly below those of some comparators, whereas Spearman correlations were marginally higher. These results demonstrate strong overall performance across both multi-protein and single-protein benchmarks, with the most pronounced gains observed on p53 and myoglobin.

We further examined whether predictive performance depended on the source of the structural inputs. Using ESMFold³⁹ instead of AlphaFold3 (AF3)⁵⁵ structures for training and validation had little effect on performance, while AF3 structures at inference yielded a modest correlation improvement of approximately 0.01 (**Supplementary Fig. 5**). These results support the applicability of STAB-DKL to structures generated by either method.

STAB-DKL improves protein-level prioritization of stabilizing mutations

Beyond predicting stability changes, a practical objective is to identify stabilizing variants among candidate mutations within a target protein. To evaluate this capability, we assembled HM7 (highly stabilizing mutations across seven proteins), a benchmark for stabilizing mutation prioritization. Protein targets were selected from the test datasets based on the availability of at least 50 experimentally characterized mutations and at least one stabilizing mutation, defined as $\Delta\Delta G < -0.5$ kcal/mol. Five of the seven targets contained five or fewer stabilizing mutations (**Fig. 6a**). We compared STAB-DKL with Stability Oracle and retrained versions of ThermoMPNN and PROSTATA. All metrics were calculated separately for each protein and then averaged across proteins. Pearson and Spearman correlations assessed overall prediction performance, while four additional metrics evaluated prioritization among the top K candidates: NDCG@K, which rewards ranking stronger stabilizing mutations nearer the top; Precision@K, the fraction of selected candidates that are stabilizing; Recall@K, the proportion of all experimentally identified stabilizing mutations recovered; and F1@K, the harmonic mean of precision and recall.

STAB-DKL achieved the highest mean Pearson and Spearman correlations on HM7, reaching 0.715 and 0.713, compared with 0.670 and 0.646 for the strongest comparators (**Fig. 6b**). Its advantage was more pronounced in ranking stabilizing mutations. Relative to ThermoMPNN, the strongest comparator for NDCG, STAB-DKL increased NDCG@5 from 0.325 to 0.546, NDCG@10 from 0.345 to 0.572, and NDCG@20 from 0.408 to 0.601, corresponding to relative improvements of 68%, 66%, and 47%, respectively (**Fig. 6c**). Precision@1, Precision@5, and Precision@10 reached 0.571, 0.371,

and 0.295, compared with the best corresponding comparator values of 0.286, 0.2, and 0.2 (**Fig. 6d**). Thus, STAB-DKL recovered a higher proportion of experimentally verified stabilizing mutations among its top-ranked candidates.

Recall@K and F1@K showed similar improvements (**Supplementary Fig. 6a-b**), and the ranking advantage was maintained under alternative stabilizing thresholds of -0.25 , -0.75 , and -1.0 kcal/mol (**Supplementary Fig. 6c**). Notably, Stability Oracle and ThermoMPNN achieved similar correlation scores but differed substantially in ranking performance, highlighting that ranking-based metrics more directly assess a model's ability to prioritize stabilizing variants than overall correlations. Incorporating predicted variance into the ranking score yielded small additional improvements in NDCG (**Fig. 6e**). Together, these results support the utility of STAB-DKL for prioritizing stabilizing mutations for experimental evaluation, with uncertainty estimates providing complementary information for candidate selection.

Discussion

In this study, we introduced STAB-DKL, an uncertainty-aware multi-task deep kernel learning framework that integrates heterogeneous stability measurements and complementary protein representations. By combining selective task sharing with dataset-specific GP modeling, STAB-DKL improved average predictive performance and protein-level prioritization of stabilizing mutations across the evaluated benchmarks.

STAB-DKL demonstrates how multi-task learning with selective representation sharing and dataset-specific probabilistic modeling can exploit complementary information across heterogeneous stability datasets. Megascale assays densely sample mutational landscapes in relatively small proteins, whereas conventional measurements cover a broader range of protein lengths. ThermoMPNN¹⁸ and Stability Oracle²⁰ established the value of megascale data but reported limited overall gains from incorporating conventional measurements under their respective training protocols. In STAB-DKL, curated and megascale $\Delta\Delta G$ measurements jointly inform shared representations, while separate GP heads accommodate dataset-specific covariance structure and observation noise. Ablation analyses of the training datasets and model architecture support both the complementary contributions of these datasets and the benefit of retaining separate GP heads. Their predictive distributions are combined through precision-weighted fusion, allowing each branch's contribution to vary across mutations

according to its estimated uncertainty.

Integrating ΔT_m extends this strategy to a related but non-equivalent measure of protein stability. SPIRED-Stab²¹ previously combined $\Delta\Delta G$ and ΔT_m supervision through shared representations and task-specific prediction heads. In STAB-DKL, ΔT_m serves as an auxiliary task that informs the shared feature transformation while retaining its own MLP and GP head. Removing Tm2403 reduced $\Delta\Delta G$ prediction performance, supporting the value of incorporating ΔT_m data. Extending MLP sharing to all three tasks also impaired $\Delta\Delta G$ prediction, suggesting that restricting parameter sharing between ΔT_m and $\Delta\Delta G$ to the LA module may limit interference with $\Delta\Delta G$ -specific learning. The gains from incorporating C4238 and Tm2403 may also reflect the larger training set and broader protein coverage they provide.

Consistent with previous studies highlighting the value of structural information and sequence-structure integration¹⁸⁻²⁰, our feature ablations show that combining pretrained sequence and structural representations substantially improves stability prediction. The additional contribution of StrucMSA, together with the limited effect of removing conventional MSA features (MSA-Seq), suggests that local structural context helps make alignment-derived evolutionary information more useful alongside pretrained embeddings. Replacing MSA features with MSA Transformer embeddings yielded comparable Pearson correlations, and combining the two did not consistently improve prediction, supporting StrucMSA as a compact alternative within this framework. Generating StrucMSA representations was also 18 to 42 times faster than generating MSA Transformer embeddings in our implementation. Together, these findings support compact, structurally informed evolutionary features as an efficient complement to pretrained sequence and structural representations.

Translating these gains in predictive accuracy into informed experimental decisions requires uncertainty estimates for individual mutations. GP-based uncertainty estimates have informed experimental selection in protein engineering and biological discovery^{30,31,56,57}, as well as confidence assessment and inverse-variance aggregation in studies of protein variation^{32,33,58,59}. Within stability prediction, BayeStab⁶⁰ incorporated Bayesian uncertainty estimation with an emphasis on separating model uncertainty from data noise. Here, STAB-DKL's dataset-specific GP modeling outperformed the tested MC dropout baseline in both predictive accuracy and uncertainty calibration. Although Gaussian NLL regression generally performed better on uncertainty metrics, its predictive accuracy

was substantially lower than that of both STAB-DKL and MC dropout. These comparisons highlight a favorable balance between predictive performance and uncertainty quantification in STAB-DKL. Furthermore, STAB-DKL showed lower errors among low-variance predictions, supporting the use of its uncertainty estimates as complementary information for assessing prediction reliability.

STAB-DKL's ranking performance on HM7 highlights its value in prioritizing stabilizing mutations for experimental testing. Previous studies, including SPIRED-Stab²¹ and DVE-stability²², assessed the overlap between predicted and experimentally determined top K mutations within each protein. Our HM7 analysis evaluates both the proportion of experimentally verified stabilizing mutations among the top K predictions and whether mutations with stronger stabilizing effects are ranked higher. This emphasis is particularly relevant given the scarcity of stabilizing mutations in most HM7 candidate pools. Compared with established predictors such as ThermoMPNN¹⁸ and Stability Oracle²⁰, STAB-DKL identified a higher proportion of stabilizing mutations among its top-ranked candidates and achieved higher NDCG. These advantages were not fully captured by conventional correlation metrics. Incorporating predictive uncertainty further improved ranking, providing complementary information for candidate selection. Together, these findings highlight the importance of assessing stability predictors by both the recovery of stabilizing mutations and the magnitude of stability gains among their top-ranked candidates.

STAB-DKL still has several limitations that need to be considered in practice and addressed in future work. The model is trained and evaluated primarily on in vitro stability measurements, which do not fully capture the cellular processes that govern protein folding, degradation and function^{1,61,62}. Integrating stability measurements with cellular phenotypes could help connect predicted thermodynamic effects to changes in protein abundance^{6,7}, aggregation^{59,63} and activity^{6,33,64-66}. Our evaluation also focuses on single substitutions. Extending the framework to combinations of mutations will require assessing epistatic interactions that can alter their combined effects⁶⁷. Finally, the ranking results derive from retrospective candidate pools spanning a limited number of proteins. Prospective testing across more diverse targets will be important for validating STAB-DKL's ability to prioritize stabilizing mutations.

In summary, STAB-DKL provides a framework for learning from complementary stability measurements while accounting for dataset differences and quantifying predictive uncertainty. Its

428 selective task sharing and dataset-specific GP heads offer a basis for incorporating additional datasets
429 with different protein coverage and experimental readouts. More broadly, STAB-DKL could help
430 explore how sequence variation shapes protein folding stability, offering insights into disease-
431 associated protein dysfunction and guiding the design of stable, functional proteins.

Methods

Dataset Curation

The megascale dataset¹⁷ was obtained from the Zenodo repository (<https://zenodo.org/records/7992926>), using the file Tsuboyama2023_Dataset2_Dataset3_20230416.csv. Only entries corresponding to natural proteins were retained, and records lacking $\Delta\Delta G$ measurements or containing more than one amino-acid substitution were excluded. The S2648⁶⁸, Q3488⁴⁸, P53⁴⁸, Myoglobin⁴⁸, S669¹⁶, and Ssym⁴² datasets were obtained from the PROSTATA²⁸ GitHub repository (<https://github.com/AIRL-Institute/PROSTATA>). The S9028, thermoMut and FireProt datasets were obtained from ddMut⁴⁹ (<https://biosig.lab.uq.edu.au/ddmut/data>), ThermoMutDB⁴⁵ (<https://biosig.lab.uq.edu.au/thermomutdb/>), and FireProtDB⁵⁰ (<https://loschmidt.chemi.muni.cz/fireprotodb/>), respectively. To construct the curated $\Delta\Delta G$ dataset, records from S2648, Q3488, S9028, thermoMut, and FireProt datasets were combined and overlapping entries were reconciled. For mutations with multiple reported values, the entry with the largest absolute $\Delta\Delta G$ was retained. For FireProt entries annotated with multiple PDBs, preference was given to PDBs already represented in other training datasets; otherwise, the first listed PDB was used.

Wild-type sequences were derived from the corresponding PDB files, and mutant sequences were generated according to the annotated substitutions, yielding 7,486 mutation records. To minimize sequence leakage into the held-out benchmark sets, sequence similarity between candidate training proteins and test proteins was assessed using MMseqs2 (-e inf --alignment-mode 3 --cov-mode 0). The resulting alignments were used to calculate a coverage-adjusted sequence identity as

$$I = \text{pident} \times \frac{L_{\text{alignment}}}{\max(L_{\text{query}}, L_{\text{target}})} \quad (1)$$

where pident is the fraction of identical residues within the aligned region, $L_{\text{alignment}}$ is the alignment length, and L_{query} and L_{target} are the lengths of the query and target sequences, respectively. Equivalently, this metric corresponds to the number of identical residues in the alignment divided by the length of the longer sequence. This coverage-adjusted measure was used to quantify sequence similarity between data partitions and identify highly similar sequences that could introduce potential information leakage. Sequences with >30% coverage-adjusted sequence identity to any test

set sequence were excluded, resulting in the curated dataset C4238. Similar procedures were applied to thermoMut and FireProt, yielding Tm2403.

For the megascale dataset, sequences with >30% coverage-adjusted sequence identity to any held-out test set sequence were first removed to minimize potential sequence-level information leakage. To construct a validation set that was sequence-separated from all training data sources, the remaining megascale sequences were pooled with sequences from C4238 and Tm2403 for clustering using MMseqs2 easy-cluster (--min-seq-id 0.3 -c 0). Megascale sequences belonging to singleton clusters in this combined sequence pool were preferentially selected until 20 proteins were reserved as the validation set (Ms20). The remaining megascale proteins were assigned to the training set (Ms154). C4238 and Tm2403 sequences were used only for clustering and were not included in either Ms20 or Ms154.

Data augmentation

We employed two previously described strategies for data augmentation. Thermodynamic reversibility (TR) leverages the state-function property of Gibbs free energy to compute $\Delta\Delta G$ for reverse mutations. For example, the $\Delta\Delta G$ of S→A is the negative of that measured for A→S, following previous implementation in Thiltgen, G. et al⁶⁹ (Supplementary Fig. 2a). Thermodynamic permutation (TP) generates additional mutation data at positions for which multiple substitutions have been experimentally characterized, as described in Daniel J. Diaz et al²⁰. For example, if the wild-type residue is S and experimental $\Delta\Delta G$ values are available for S→G and S→A mutations, the $\Delta\Delta G$ for G→A can be inferred from the difference between the two measurements, $\Delta\Delta G_{S\rightarrow A} - \Delta\Delta G_{S\rightarrow G}$, with A→G given by the negative of this value (Supplementary Fig. 2a).

Data Sampling

Given the large size of the augmented megascale dataset and the potential redundancy in information contributed by multiple variants from the same protein, we applied proportional, sequence-based sampling to the augmented Ms154 dataset. Specifically, sampling was performed while preserving the original distribution of mutation counts across distinct protein sequences. We evaluated subsets containing 10,000, 20,000, 40,000 and 60,000 samples. Model performance approached a

plateau at approximately 20,000 samples, with only modest additional gains at larger sample sizes. We therefore used 60,000 samples for the final model to leverage the available training data while retaining these incremental performance gains. For the validation dataset (Ms20), we applied the same proportional, sequence-based sampling strategy to select 500 forward-mutation records. Corresponding reverse mutations were then generated using TR, yielding separate forward- and reverse-mutation validation sets of 500 records each. Both sets were used for model selection.

Training label scaling

Before training, the megascale $\Delta\Delta G$ values and ΔT_m values were rescaled to match the standard deviation of the curated $\Delta\Delta G$ training values. For each dataset, values were multiplied by the ratio of the curated $\Delta\Delta G$ standard deviation to that dataset's standard deviation. Standard deviations were calculated from the training data after augmentation and sampling.

STAB-DKL Architecture

STAB-DKL is a multi-task deep kernel learning framework consisting of two main components: a multi-modal feature extractor and a multi-task deep kernel learner. The framework integrates sequence, structural and structure-constrained evolutionary representations to jointly model $\Delta\Delta G$ and ΔT_m , with ΔT_m serving as an auxiliary task during training.

Feature reuse during data augmentation

Structural and StrucMSA features were generated once from the original WT sequences and reused for the corresponding augmented variants, with mutation-specific differences recalculated for each substitution. The MSA features depend on the set of homologous sequences identified in the initial search, which is expected to remain largely unchanged after a single amino-acid substitution. Structural feature reuse avoids redundant structure prediction across the large number of augmented variants.

Sequence features

For each mutation, we first obtained the 33rd-layer hidden representations for both the wild-type and mutant sequences using the pretrained ESM2-650M model³⁹. The residue-level embeddings corresponding to the mutation site were then extracted, and their difference was used as the sequence representation, reflecting the local perturbation induced by the mutation in the learned sequence space.

Structural features

ProteinMPNN⁴⁰ was used to derive structure-informed representations. WT structures were processed under two decoding settings—with and without sequence conditioning. From both decoding modes, residue-level embeddings and hidden-state vectors at the mutation position were collected. Additionally, we extracted the log probabilities of 20 amino acids at the mutation site from ProteinMPNN’s final layer and computed the difference between WT and mutant log probabilities. These components were concatenated to form the final structural feature vector.

StrucMSA features

To incorporate evolutionary information while respecting structural context, we generated structure-constrained MSAs using MMseqs2 (UniRef90 database). Each MSA was filtered according to the local structural proximity to the mutation site. Specifically, the distance of each other residue was defined as the minimum distance between the C α atom of the mutated residue and any atom of that residue. For each radius from 1 to 20 Å, at 1 Å intervals, homologous sequences were retained if they matched the wild-type sequence at $\geq 50\%$, $\geq 75\%$ or 100% of the neighboring positions within that radius. The 60 filtering conditions, together with the original unfiltered alignment, yielded 61 amino-acid frequency profiles. For each alignment, the log-probability difference between wild-type and mutant residues at the mutation site was computed and concatenated to form the final MSA feature representation.

Multi-task deep kernel learning module

Following feature extraction, the concatenated input features for all tasks were passed through a shared LA module. The curated and megascale $\Delta\Delta G$ tasks shared the same multilayer perceptron (MLP; hidden sizes: 1024 and 256; output size: 2), whereas the ΔT_m task used a separate MLP with the same architecture. Each task was modeled by a separate GP using a radial basis function (RBF) kernel with automatic relevance determination (ARD), together with a Gaussian likelihood. To enable scalable GP inference, the GP covariance functions were approximated using KISS-GP with grid interpolation. The GP mean function, kernel hyperparameters and observation-noise parameters were task-specific and optimized by minimizing the weighted negative marginal log likelihood. Each GP yielded a predictive Gaussian distribution characterized by a predictive mean and variance.

The two $\Delta\Delta G$ predictions were subsequently combined by precision-weighted fusion. Let μ_i and σ_i^2 denote the predictive mean and variance of the i -th $\Delta\Delta G$ task, respectively. The fused predictive mean is given by:

$$\hat{y} = \frac{\sum_i (\mu_i / \sigma_i^2)}{\sum_i (1 / \sigma_i^2)} \quad (2)$$

and the corresponding predictive variance is:

$$\widehat{\sigma^2} = \frac{1}{\sum_i (1 / \sigma_i^2)} \quad (3)$$

where the predictive variances used for fusion were obtained after applying the Gaussian likelihood. This precision-weighted scheme assigns greater weight to predictions with lower uncertainty while accounting for the task-specific predictive variance.

Training Configuration

STAB-DKL was trained using negative Exact Marginal Log Likelihood (EMLL) as the objective, with the total loss computed as a weighted sum of three tasks— ΔT_m , curated $\Delta\Delta G$, and megascale $\Delta\Delta G$ —with weights of 0.5, 0.25 and 0.25 respectively. The higher weight for ΔT_m reflects its role as auxiliary supervision, while the curated and megascale $\Delta\Delta G$ tasks are assigned equal weights in the training objective.

Optimization was performed using the Adam optimizer (betas = (0.9, 0.99)), with a learning rate of 1×10^{-2} for the GP layers and a learning rate sweep from 1×10^{-4} to 1×10^{-2} for the LA module and MLP layers. A Cosine Annealing LR schedule was employed for the learning rate. Models were trained for 50 epochs, with the optimal model selected based on validation performance.

For validation, the Ms20 dataset was augmented via thermodynamic reversibility (TR), producing two validation sets containing forward and reverse mutations, respectively. This ensured that model evaluation accounted for both mutation directions. Among epochs ≥ 30 , the best model was selected based on the maximum sum of Pearson and Spearman correlation coefficients across the two validation sets.

Post-training prediction scaling

To calibrate the magnitude of the predicted $\Delta\Delta G$ values, post hoc scaling was performed using the validation set. A scaling factor s was initialized at 1.0 and increased iteratively by 0.1. At each step, the raw predictions were divided by the corresponding scaling factor:

$$\hat{y}_{\text{scaled}} = \frac{\hat{y}_{\text{raw}}}{s} \quad (4)$$

where $\hat{y}_{\text{raw}}$ and $\hat{y}_{\text{scaled}}$ denote the raw and scaled $\Delta\Delta G$ predictions, respectively. The selected factor was updated only when a candidate reduced validation MSE by more than 10% relative to the most recently accepted factor. The resulting scaling factor was then fixed and applied to the final predictions.

Monte Carlo dropout

Monte Carlo dropout is a technique for uncertainty estimation in neural network predictions, in which dropout is interpreted as approximate Bayesian inference. During inference, dropout layers remain active and multiple stochastic forward passes are performed to sample from an approximate predictive distribution. In our study, dropout was enabled during inference and 50 stochastic forward passes were performed for each input. The predictive mean and variance were calculated from the resulting predictions.

Gaussian negative log-likelihood

The Gaussian negative log-likelihood (NLL) is a probabilistic loss function commonly used in regression tasks to jointly learn predictive means and variances. It assumes that each target value y is drawn from a Gaussian distribution parameterized by the model's predicted mean μ and variance σ^2 .

In our application, we optimized the model by minimizing the Gaussian NLL, which penalizes both large residual errors and overconfident uncertainty estimates. The loss function is formulated as:

$$\mathcal{L}_{\text{NLL}} = \frac{1}{2} \left[\log(2\pi\sigma^2) + \frac{(y-\mu)^2}{\sigma^2} \right] \quad (5)$$

where μ and σ^2 represent the predicted mean and variance, respectively, and y is the true target value.

Uncertainty evaluation

All uncertainty analyses were performed on the final fused $\Delta\Delta G$ predictions and their corresponding predictive variances. Calibration was evaluated using the miscalibration area (MA), calculated across nominal confidence levels from 0.01 to 0.99 at 0.01 intervals. For each confidence level (c), the prediction interval was defined as $\hat{y} \pm z_c \sigma$, where $z_c = \Phi^{-1}[(1+c)/2]$, and empirical coverage was calculated as the proportion of test samples whose observed values fell within the corresponding intervals. MA was calculated as the mean absolute difference between empirical and

nominal coverage across all confidence levels:

$$\frac{1}{99} \sum_{j=1}^{99} |\text{Coverage}(c_j) - c_j| \quad (6)$$

Predictive likelihood was evaluated using the mean Gaussian negative log likelihood (NLL):

$$\frac{1}{N} \sum_{i=1}^N \left[\frac{(y_i - \mu_i)^2}{2\sigma_i^2} + \frac{1}{2} \log(2\pi\sigma_i^2) \right] \quad (7)$$

where μ_i and σ_i^2 denote the predicted mean and variance, respectively. The association between predictive uncertainty and prediction error was further assessed using Spearman correlation between the predictive variance and the absolute residual, $|y_i - \hat{y}_i|$. For selective prediction analysis, samples were ranked by predictive variance, and subsets with the lowest variance were retained at predefined retention fractions. All metrics were calculated using the final fused $\Delta\Delta G$ prediction.

Evaluation metrics

We evaluated STAB-DKL using two classes of metrics to capture different aspects of model performance: regression-based and ranking-based metrics.

For regression performance, we reported the Pearson and Spearman correlation coefficients, as well as root mean squared error (RMSE) and mean absolute error (MAE).

To assess the model's ability to prioritize stabilizing mutations, we used a modified NDCG@K metric. For a given set of predicted mutations, let y_i denote the experimental $\Delta\Delta G$ of mutation i (more negative $\Delta\Delta G$ indicates a greater stabilizing effect). Mutations with $y_i < -0.5$ kcal/mol were considered stabilizing.

Discounted cumulative gain (DCG@K) is defined as:

$$\text{DCG@K} = \sum_{i=1}^K \frac{y_i \cdot \mathbf{1}(y_i < -0.5)}{\log_2(r_i + 1)} \quad (8)$$

where r_i is the rank of mutation i in the predicted top K (ordered from most negative predicted $\Delta\Delta G$ to less negative), and $\mathbf{1}(\cdot)$ is the indicator function selecting only stabilizing mutations. Ideal DCG (IDCG@K) is computed analogously from the experimentally observed $\Delta\Delta G$ values, using the top K most stabilizing mutations:

$$\text{IDCG@K} = \sum_{i=1}^K \frac{y_i^{\text{ideal}} \cdot \mathbf{1}(y_i^{\text{ideal}} < -0.5)}{\log_2(r_i^{\text{ideal}} + 1)} \quad (9)$$

Finally, the normalized DCG is:

$$\text{NDCG@K} = \frac{\text{DCG@K}}{\text{IDCG@K}} \quad (10)$$

This metric emphasizes the model's ability to rank the most stabilizing mutations at the top of the prediction list. Precision@K, Recall@K and F1@K were additionally computed to quantify the fraction of stabilizing mutations captured in the top K predictions:

$$\text{Precision@K} = \frac{\# \text{ of stabilizing mutations in top K}}{K} \quad (11)$$

$$\text{Recall@K} = \frac{\# \text{ of stabilizing mutations in top K}}{\text{total \# of stabilizing mutations}} \quad (12)$$

$$\text{F1@K} = 2 \times (\text{Precision@K} \times \text{Recall@K}) \div (\text{Precision@K} + \text{Recall@K}) \quad (13)$$

Uncertainty-guided mutation prioritization

For each protein, we used the model's predicted mean and variance to compute an uncertainty-adjusted prioritization score for all target mutations. Specifically, the predictive variance was first shifted to impose a positive lower bound:

$$\text{var}' = \text{var} - \min(\text{var}) + \lambda \quad (14)$$

where λ is a variance adjustment term that defines the lower bound of the effective variance and modulates the contribution of predictive uncertainty to the prioritization score. Next, a probability-based stability score was computed as:

$$\text{score} = -\Phi(-0.5; \mu, \sqrt{\text{var}'}) \quad (15)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a Gaussian distribution with predictive mean μ and variance var' . The resulting score incorporates both the predicted $\Delta\Delta G$ and its associated uncertainty, with lower scores corresponding to mutations with a higher probability of satisfying the stability criterion ($\Delta\Delta G < -0.5$).

PROSTATA retraining

The PROSTATA model was reproduced using the official source code available at <https://github.com/AIRI-Institute/PROSTATA>. The model was retrained using the megascale $\Delta\Delta G$ training set used in our study, without a separate validation set. Following the original implementation, the final prediction was obtained by averaging the outputs from five separately trained models with

different regression heads. Training was performed with a batch size of 8 and gradient accumulation over 8 steps. The model was optimized using the Adam optimizer with a learning rate of 1×10^{-5} for 3 epochs.

ThermoMPNN retraining

The ThermoMPNN model was obtained from the official repository at <https://github.com/Kuhlman-Lab/ThermoMPNN>. The model was retrained using the same megascale $\Delta\Delta G$ training-validation split used in our study. No thermodynamic reversibility or permutation augmentation was applied to the training data. All other hyperparameters followed the original ThermoMPNN implementation.

Data and code availability: The STAB-DKL source code, documentation, example files, and processed training, validation and test datasets are available on GitHub at <https://github.com/wanglab-szbl/STAB-DKL>. Trained model weights and precomputed features are available from Zenodo at <https://doi.org/10.5281/zenodo.20339667>.

Acknowledgements: We acknowledge the High Performance Computing and Informatics Core Facility of Shenzhen Bay Laboratory and the Computing Labware for Electron Microscopy Visualization and Experimental Research at the Shenzhen Medical Academy of Research and Translation (SMART) for providing access to computational infrastructure. Grants were provided by National Key R&D Program of China (2025YFA0923000 to C.W.), Guangdong Provincial Key R&D Program (2024B1111150001 to C.W.), National Natural Science Foundation of China (82471903 to C.W.), Guangdong Pearl River Funding (2023QN10Y299 to C.W.), and Shenzhen Bay Laboratory (21240111 to C.W.).

Author contributions: C.W. conceived the project. C.W. and J.H. designed the study. J.H. performed the computational modeling and analysis. J.H. and C.W. wrote the manuscript.

Competing interests: Shenzhen Bay Laboratory has filed a patent application related to the method described in this work with the China National Intellectual Property Administration (application no.

202511863717; C.W. is a named inventor). The authors declare no other competing interests.

References

- 1 Balchin, D., Hayer-Hartl, M. & Hartl, F. U. In vivo aspects of protein folding and quality control. *Science* **353**, aac4354 (2016).
- 2 Gershenson, A., Gierasch, L. M., Pastore, A. & Radford, S. E. Energy landscapes of functional proteins are inherently risky. *Nat Chem Biol* **10**, 884-891 (2014).
- 3 Sumida, K. H. *et al.* Improving Protein Expression, Stability, and Function with ProteinMPNN. *J Am Chem Soc* **146**, 2054-2061 (2024).
- 4 Wilson, A. E., Kosater, W. M. & Liberles, D. A. Evolutionary Processes and Biophysical Mechanisms: Revisiting Why Evolved Proteins Are Marginally Stable. *J Mol Evol* **88**, 415-417 (2020).
- 5 Stein, A., Fowler, D. M., Hartmann-Petersen, R. & Lindorff-Larsen, K. Biophysical and Mechanistic Models for Disease-Causing Protein Variants. *Trends in Biochemical Sciences* **44**, 575-588 (2019).
- 6 Cagiada, M. *et al.* Understanding the Origins of Loss of Protein Function by Analyzing the Effects of Thousands of Variants on Activity and Abundance. *Mol Biol Evol* **38**, 3235-3246 (2021).
- 7 Beltran, A., Jiang, X., Shen, Y. & Lehner, B. Site-saturation mutagenesis of 500 human protein domains. *Nature* **637**, 885-894 (2025).
- 8 Høie, M. H., Cagiada, M., Beck Frederiksen, A. H., Stein, A. & Lindorff-Larsen, K. Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell Rep* **38**, 110207 (2022).
- 9 Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology* **10**, 866-876 (2009).
- 10 Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology* **19**, 596-604 (2009).
- 11 Goldenzweig, A. & Fleishman, S. J. Principles of Protein Stability and Their Application in Computational Design. *Annu Rev Biochem* **87**, 105-129 (2018).
- 12 Listov, D., Goverde, C. A., Correia, B. E. & Fleishman, S. J. Opportunities and challenges in design and optimization of protein function. *Nat Rev Mol Cell Biol* **25**, 639-653 (2024).
- 13 Marabotti, A., Scafuri, B. & Facchiano, A. Predicting the stability of mutant proteins by computational approaches: an overview. *Briefings in Bioinformatics* **22**, bbaa074 (2021).
- 14 Leman, J. K. *et al.* Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nature Methods* **17**, 665-680 (2020).
- 15 Delgado, J., Radusky, L. G., Cianferoni, D. & Serrano, L. FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics* **35**, 4168-4169 (2019).
- 16 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** (2022).
- 17 Tsuboyama, K. *et al.* Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434-444 (2023).
- 18 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 U S A*

- 736 **121**, e2314853121 (2024).
- 737 19 Li, Z. & Luo, Y. Generalizable and scalable protein stability prediction with rewired protein
738 generative models. *Nat Commun* **17**, 891 (2026).
- 739 20 Diaz, D. J. *et al.* Stability Oracle: a structure-based graph-transformer framework for
740 identifying stabilizing mutations. *Nat Commun* **15**, 6170 (2024).
- 741 21 Chen, Y., Xu, Y., Liu, D., Xing, Y. & Gong, H. An end-to-end framework for the prediction of
742 protein structure and fitness from single sequence. *Nat Commun* **15**, 7400 (2024).
- 743 22 Nie, Z. *et al.* Predicting protein stability changes upon mutations with dual-view ensemble
744 learning from single sequence. *Briefings in Bioinformatics* **26**, bbaf319 (2025).
- 745 23 Montanucci, L. *et al.* DDGun: an untrained predictor of protein stability changes upon amino
746 acid variants. *Nucleic Acids Res* **50**, W222-w227 (2022).
- 747 24 Capriotti, E., Fariselli, P. & Casadio, R. I-Mutant2.0: predicting stability changes upon
748 mutation from the protein sequence or structure. *Nucleic Acids Res* **33**, W306-310 (2005).
- 749 25 Fariselli, P., Martelli, P. L., Savojardo, C. & Casadio, R. INPS: predicting the impact of non-
750 synonymous variations on protein stability from sequence. *Bioinformatics* **31**, 2816-2821
751 (2015).
- 752 26 Folkman, L., Stantic, B., Sattar, A. & Zhou, Y. EASE-MM: Sequence-Based Prediction of
753 Mutation-Induced Stability Changes with Feature-Based Multiple Models. *Journal of*
754 *Molecular Biology* **428**, 1394-1405 (2016).
- 755 27 Blaabjerg, L. M. *et al.* Rapid protein stability prediction using deep learning representations.
756 *Elife* **12** (2023).
- 757 28 Umerenkov, D. *et al.* PROSTATA: a framework for protein stability assessment using
758 transformers. *Bioinformatics* **39** (2023).
- 759 29 Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning*. (The
760 MIT Press, 2005).
- 761 30 Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with
762 Gaussian processes. *Proc Natl Acad Sci U S A* **110**, E193-201 (2013).
- 763 31 Hie, B., Bryson, B. D. & Berger, B. Leveraging Uncertainty in Machine Learning Accelerates
764 Biological Discovery and Design. *Cell Syst* **11**, 461-477 e469 (2020).
- 765 32 Wang, C. & Balch, W. E. Bridging Genomics to Phenomics at Atomic Resolution through
766 Variation Spatial Profiling. *Cell Rep* **24**, 2013-2028 e2016 (2018).
- 767 33 Zhao, P., Wang, C., Sun, S., Wang, X. & Balch, W. E. Tracing genetic diversity captures the
768 molecular basis of misfolding disease. *Nat Commun* **15**, 3333 (2024).
- 769 34 Jokinen, E., Heinonen, M. & Lähdesmäki, H. mGPfusion: predicting protein stability changes
770 with Gaussian process kernel learning and data fusion. *Bioinformatics* **34**, i274-i283 (2018).
- 771 35 Wilson, A. G., Hu, Z., Salakhutdinov, R. & Xing, E. P. *Deep Kernel Learning*. in *Proceedings*
772 *of the 19th International Conference on Artificial Intelligence and Statistics* Vol. 51 370--378
773 (PMLR, Proceedings of Machine Learning Research, 2016).
- 774 36 Wilson, A. & Nickisch, H. *Kernel Interpolation for Scalable Structured Gaussian Processes*
775 *(KISS-GP)*. in *Proceedings of the 32nd International Conference on Machine Learning* Vol. 37
776 1775--1784 (PMLR, Proceedings of Machine Learning Research, 2015).
- 777 37 Bonilla, E. V., Chai, K. M. A. & Williams, C. K. I. *Multi-task Gaussian Process prediction*. in
778 *Proceedings of the 21st International Conference on Neural Information Processing Systems*
779 153–160 (Curran Associates Inc., Vancouver, British Columbia, Canada, 2007).

- 780 38 Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein
781 engineering. *Nat Methods* **16**, 687-694 (2019).
- 782 39 Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language
783 model. *Science* **379**, 1123-1130 (2023).
- 784 40 Dauparas, J. *et al.* Robust deep learning-based protein sequence design using ProteinMPNN.
785 *Science* **378**, 49-56 (2022).
- 786 41 Stark, H., Dallago, C., Heinzinger, M. & Rost, B. Light attention predicts protein location from
787 the language of life. *Bioinform Adv* **1**, vbab035 (2021).
- 788 42 Pucci, F., Bernaerts, K. V., Kwasigroch, J. M. & Rooman, M. Quantification of biases in
789 predictions of protein stability changes upon mutations. *Bioinformatics* **34**, 3659-3665 (2018).
- 790 43 Kepp, K. P. Towards a "Golden Standard" for computing globin stability: Stability and structure
791 sensitivity of myoglobin mutants. *Biochim Biophys Acta* **1854**, 1239-1248 (2015).
- 792 44 Pires, D. E., Ascher, D. B. & Blundell, T. L. mCSM: predicting the effects of mutations in
793 proteins using graph-based signatures. *Bioinformatics* **30**, 335-342 (2014).
- 794 45 Xavier, J. S. *et al.* ThermoMutDB: a thermodynamic database for missense mutations. *Nucleic
795 Acids Res* **49**, D475-D479 (2021).
- 796 46 Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K. & Sarai, A. ProTherm, version 4.0:
797 thermodynamic database for proteins and mutants. *Nucleic Acids Res* **32**, D120-121 (2004).
- 798 47 Dehouck, Y. *et al.* Fast and accurate predictions of protein stability changes upon mutations
799 using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics* **25**, 2537-2543
800 (2009).
- 801 48 Li, B., Yang, Y. T., Capra, J. A. & Gerstein, M. B. Predicting changes in protein thermodynamic
802 stability upon point mutation with deep 3D convolutional neural networks. *PLoS Comput Biol*
803 **16**, e1008291 (2020).
- 804 49 Zhou, Y., Pan, Q., Pires, D. E. V., Rodrigues, C. H. M. & Ascher, D. B. DDMut: predicting
805 effects of mutations on protein stability using deep learning. *Nucleic Acids Res* **51**, W122-
806 W128 (2023).
- 807 50 Stourac, J. *et al.* FireProtDB: database of manually curated protein stability data. *Nucleic Acids
808 Res* **49**, D319-D324 (2021).
- 809 51 Chen, Y. *et al.* PremPS: Predicting the impact of missense mutations on protein stability. *PLoS
810 Comput Biol* **16**, e1008543 (2020).
- 811 52 Gal, Y. & Ghahramani, Z. *Dropout as a Bayesian Approximation: Representing Model
812 Uncertainty in Deep Learning.* in *Proceedings of The 33rd International Conference on
813 Machine Learning* Vol. 48 1050--1059 (PMLR, Proceedings of Machine Learning Research,
814 2016).
- 815 53 Nix, D. A. & Weigend, A. S. *Estimating the mean and variance of the target probability
816 distribution.* in *Proceedings of 1994 IEEE International Conference on Neural Networks
817 (ICNN'94).* 55-60. (1994).
- 818 54 Kendall, A. & Gal, Y. *What Uncertainties Do We Need in Bayesian Deep Learning for
819 Computer Vision?* in *Advances in Neural Information Processing Systems.* 5574-5584 (Curran
820 Associates, Inc.). (2017).
- 821 55 Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold
822 3. *Nature* **630**, 493-500 (2024).
- 823 56 Yang, J., Li, F. Z. & Arnold, F. H. Opportunities and Challenges for Machine Learning-Assisted

Enzyme Engineering. *ACS Cent Sci* **10**, 226-241 (2024).

57 Yang, J. *et al.* Active learning-assisted directed evolution. *Nat Commun* **16**, 714 (2025).

58 Wang, C., Angles, F. & Balch, W. E. Triangulating variation in the population to define mechanisms for precision management of genetic disease. *Structure* **30**, 1190-1207 e1195 (2022).

59 Chen, S., Deng, Y., Zhang, X. & Wang, C. Mapping the polymerization landscape of human serpins across genetic variation using Gaussian Processes. *LangTaoSha Preprint Server* **10.65215/LTSpreprints.2026.08.25.000315** (2026).

60 Wang, S., Tang, H., Zhao, Y. & Zuo, L. BayeStab: Predicting effects of mutations on protein stability with uncertainty quantification. *Protein Sci* **31**, e4467 (2022).

61 Hipp, M. S. & Hartl, F. U. Interplay of Proteostasis Capacity and Protein Aggregation: Implications for Cellular Function and Disease. *J Mol Biol* **436**, 168615 (2024).

62 Balch, W. E., Morimoto, R. I., Dillin, A. & Kelly, J. W. Adapting proteostasis for disease intervention. *Science* **319**, 916-919 (2008).

63 Thompson, M. *et al.* Massive experimental quantification allows interpretable deep learning of protein aggregation. *Science Advances* **11**, eadt5111 (2025).

64 Vanella, R. *et al.* Understanding activity-stability tradeoffs in biocatalysts by enzyme proximity sequencing. *Nature Communications* **15**, 1807 (2024).

65 Sun, S. *et al.* Spatial covariance reveals isothiocyanate natural products adjust redox stress to restore function in alpha-1-antitrypsin deficiency. *Cell Rep Med* **6**, 101917 (2025).

66 Sun, S. *et al.* Capturing the conversion of the pathogenic alpha-1-antitrypsin fold by ATF6 enhanced proteostasis. *Cell Chem Biol* **30**, 22-42 e25 (2023).

67 Dieckhaus, H. & Kuhlman, B. Protein stability models fail to capture epistatic interactions of double point mutations. *Protein Sci* **34**, e70003 (2025).

68 Dehouck, Y., Kwasigroch, J. M., Gilis, D. & Roman, M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics* **12**, 151 (2011).

69 Thiltgen, G. & Goldstein, R. A. Assessing predictors of changes in protein stability upon mutation using self-consistency. *PLoS One* **7**, e46084 (2012).

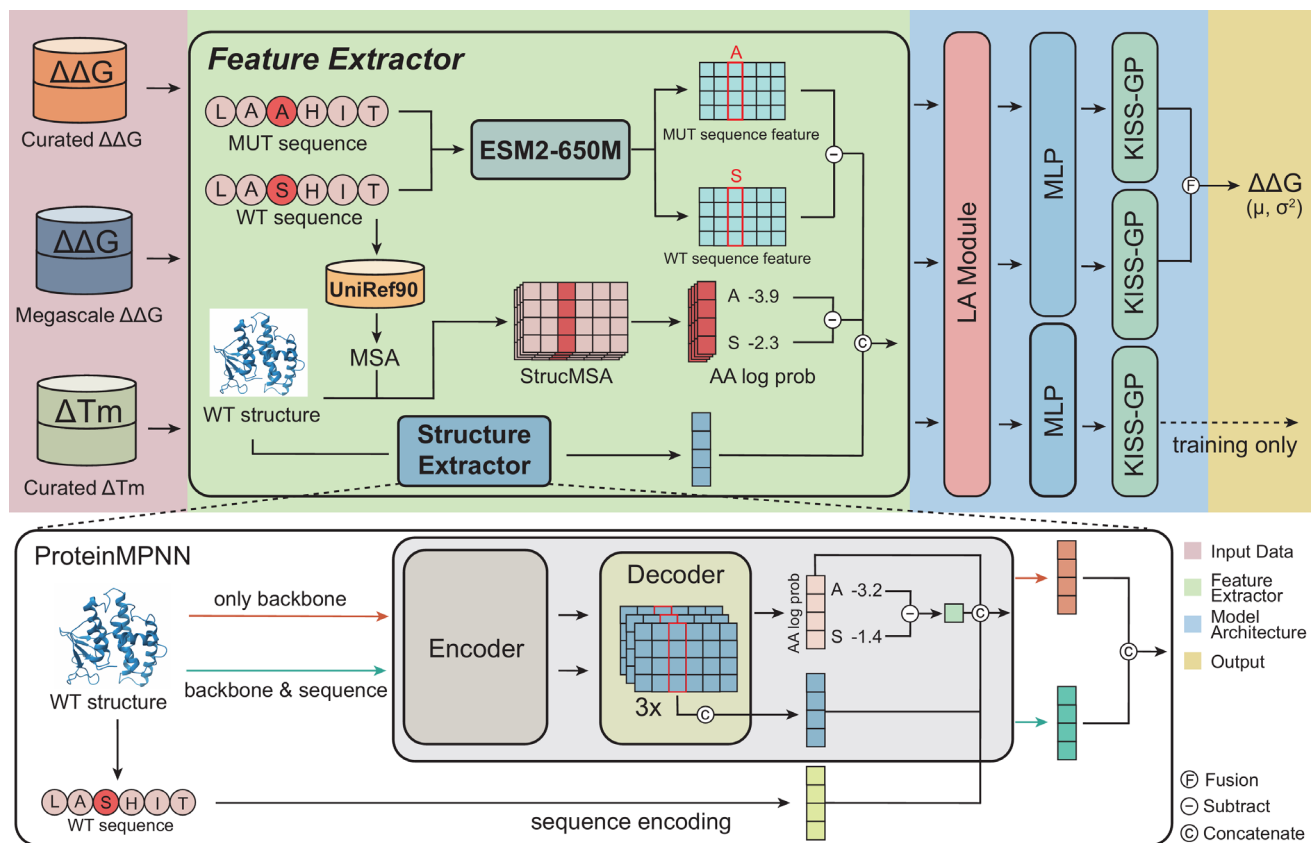


Fig. 1. Overview of STAB-DKL. Schematic of the STAB-DKL architecture illustrating how data from different sources are processed for feature extraction, training, and prediction. Features for mutations in the curated $\Delta\Delta G$ (C4238), megascale $\Delta\Delta G$ (Ms154) and curated ΔT_m (Tm2403) datasets are first obtained using the Feature Extractor. All the tasks share the same LA module; C4238 and Ms154 then use a shared MLP, whereas Tm2403 uses a separate MLP. Each branch produces a predictive mean and variance through a dataset-specific KISS-GP head. During training, the combined loss from all three tasks is used for optimization, whereas during inference, only the C4238 and Ms154 outputs are combined through precision-weighted fusion to form the final predictive distribution. In the Structure Extractor, “backbone only” means that only the protein backbone information is used during decoding, whereas “backbone & sequence” means that both backbone and sequence information are utilized.

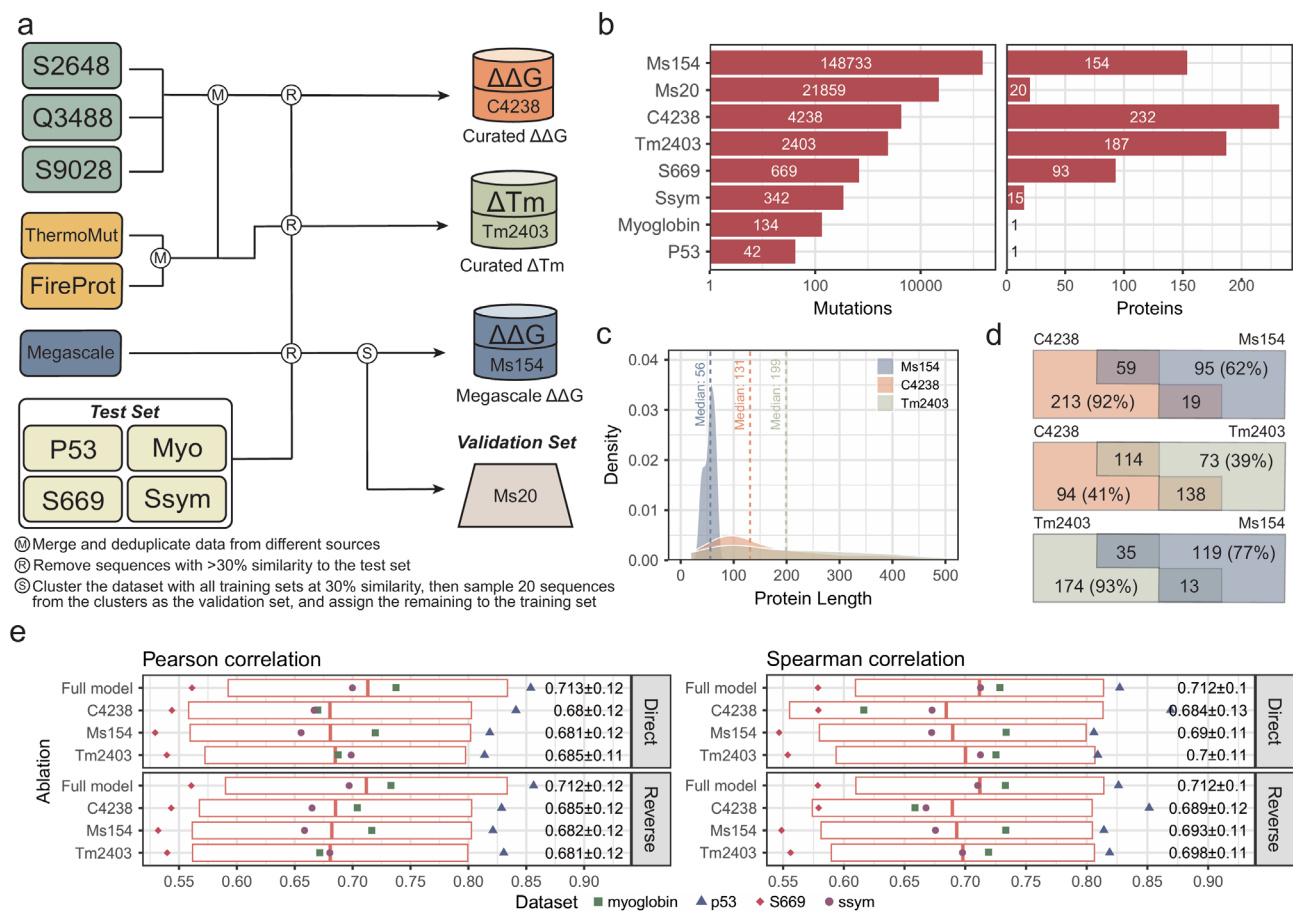


Fig. 2. Dataset curation, characterization and ablation analysis. **a**, Pipeline illustrating the processing of raw datasets into training, validation, and test sets. **b**, Bar chart depicting the numbers of proteins and mutations in each dataset. **c**, Density plots showing the distribution of protein sequence lengths in the Ms154, C4238, and Tm2403 datasets. **d**, Diagram of sequence similarity among the Ms154, C4238, and Tm2403 datasets. Numbers in the larger regions indicate the count of sequences in one dataset sharing <30% identity with any sequence in the other dataset. Numbers in the smaller regions indicate the count of sequences in one dataset sharing >30% identity with sequences in the corresponding dataset on the opposite side. **e**, Model performance in dataset ablation experiments, shown as point plots with mean $\pm$ SD.

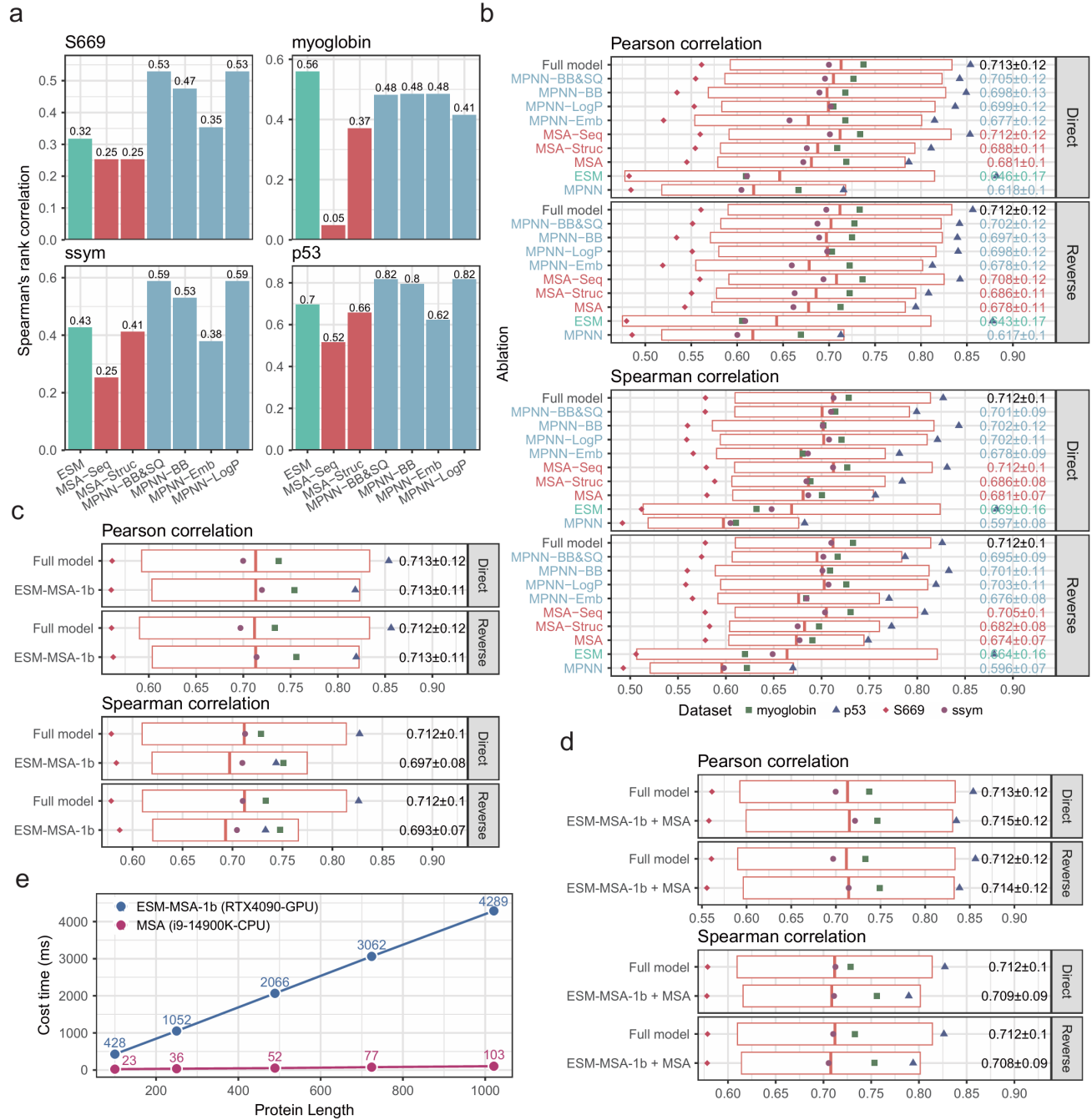


Fig. 3. Systematic assessment of feature representations and their impact on model performance.
a, Bar chart showing the absolute Spearman correlations between feature values and experimental $\Delta\Delta G$ values across different feature types in each test dataset. **b**, Model performance after ablating different feature types, shown as point plots with mean $\pm$ SD. **c**, Model performance when the original MSA features were replaced with ESM-MSA-1b embeddings, shown as point plots with mean $\pm$ SD. **d**, Model performance when the MSA features and the ESM-MSA-1b embeddings were incorporated together, shown as point plots with mean $\pm$ SD. **e**, Line plot illustrating the time required to extract different feature types as a function of protein sequence length.

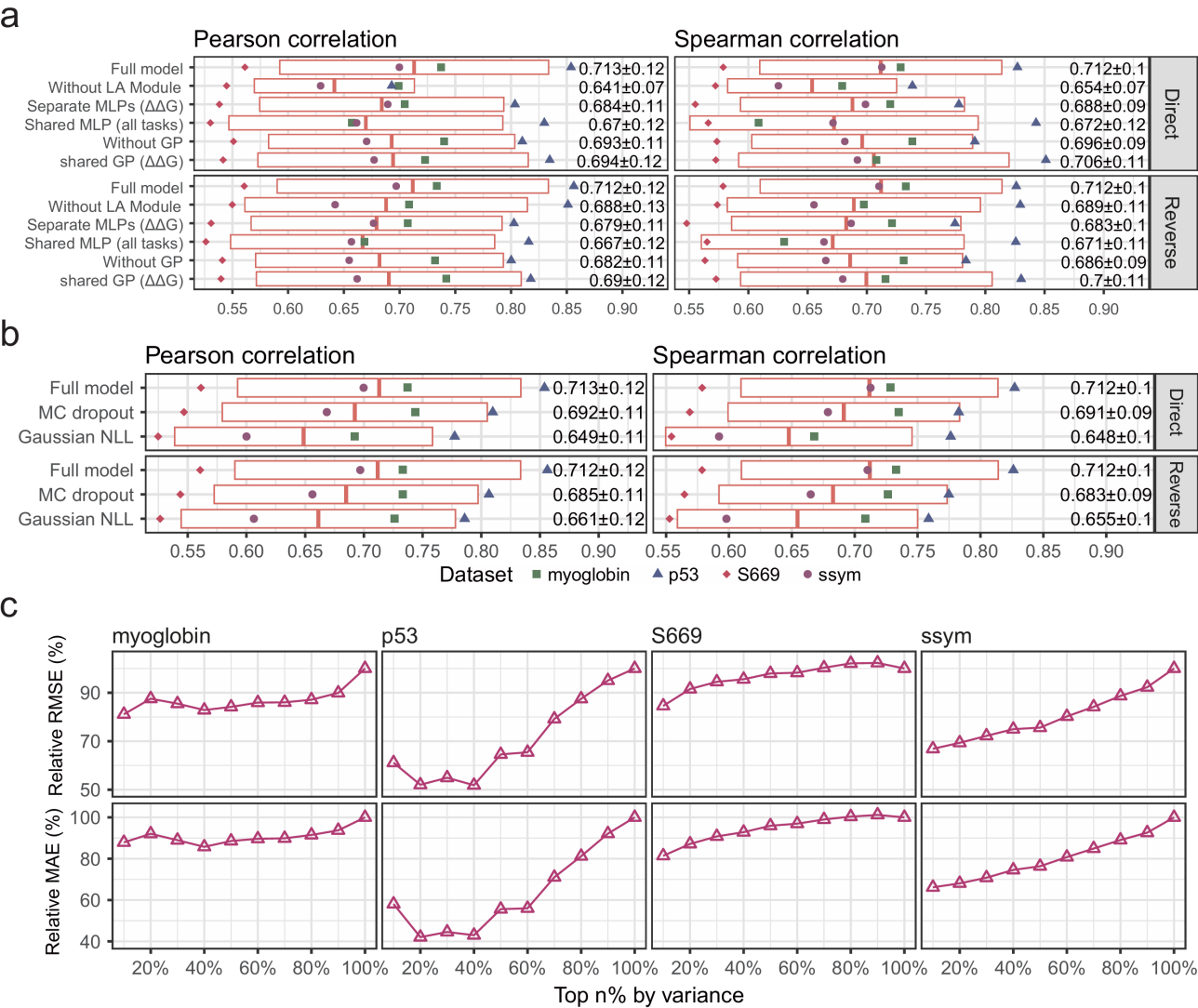


Fig. 4. Evaluation of model architecture and uncertainty assessment. a, Ablation analysis of model architecture and task-sharing strategies, shown as point plots with mean ± SD. **b**, Comparison of predictive performance among STAB-DKL, MC dropout and Gaussian NLL regression. Results are shown as point plots with mean ± SD. **c**, Relative RMSE and MAE on the four test datasets for the top n% of predictions ranked in ascending order of predictive variance.

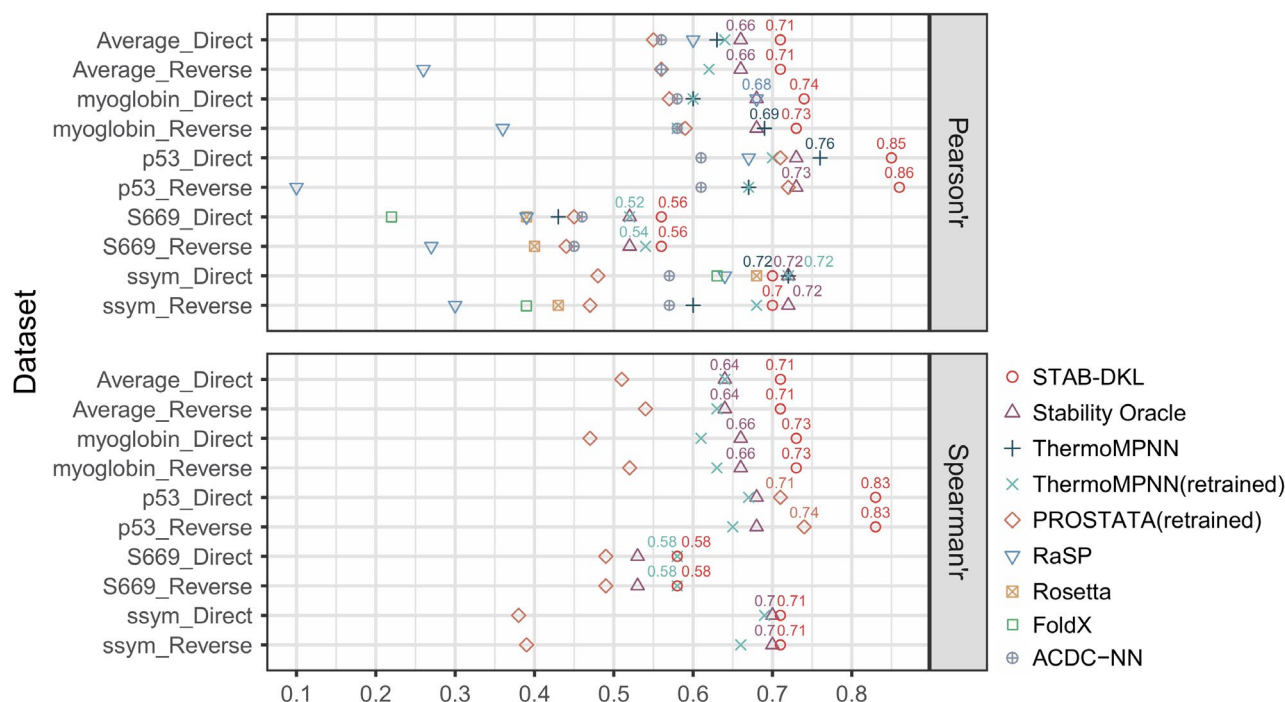


Fig. 5. Performance comparison of STAB-DKL and other methods for predicting mutation-induced stability changes. Dot plots showing Pearson (top) and Spearman (bottom) correlation coefficients for direct and reverse mutations in the myoglobin, p53, S669 and Ssym test datasets. Different colors and symbols represent different prediction methods. “Average_Direct” and “Average_Reverse” indicate the mean correlations across the four test datasets for direct and reverse mutations, respectively. Results are shown for both the original and retrained ThermoMPNN models, together with retrained PROSTATA.

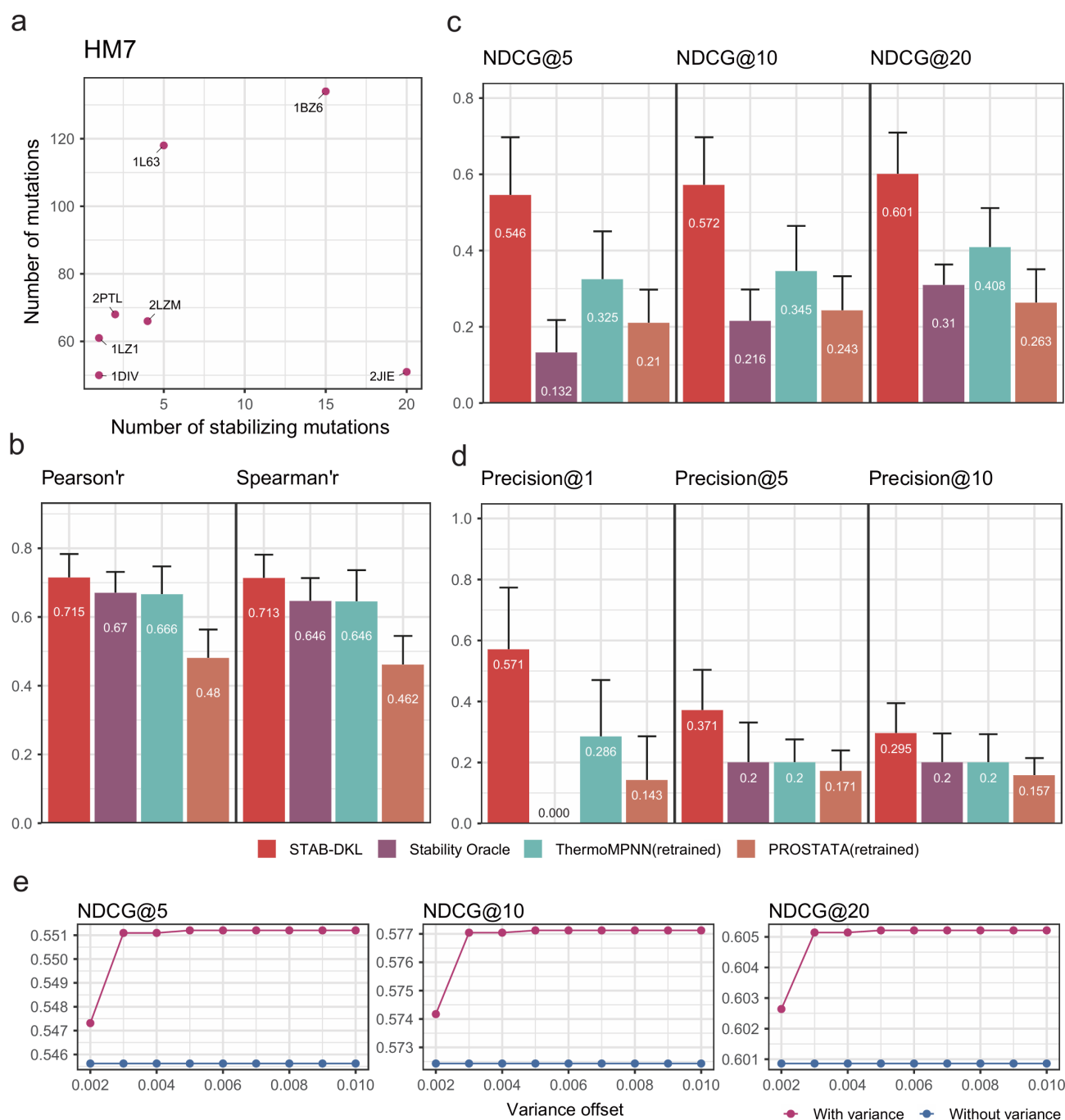
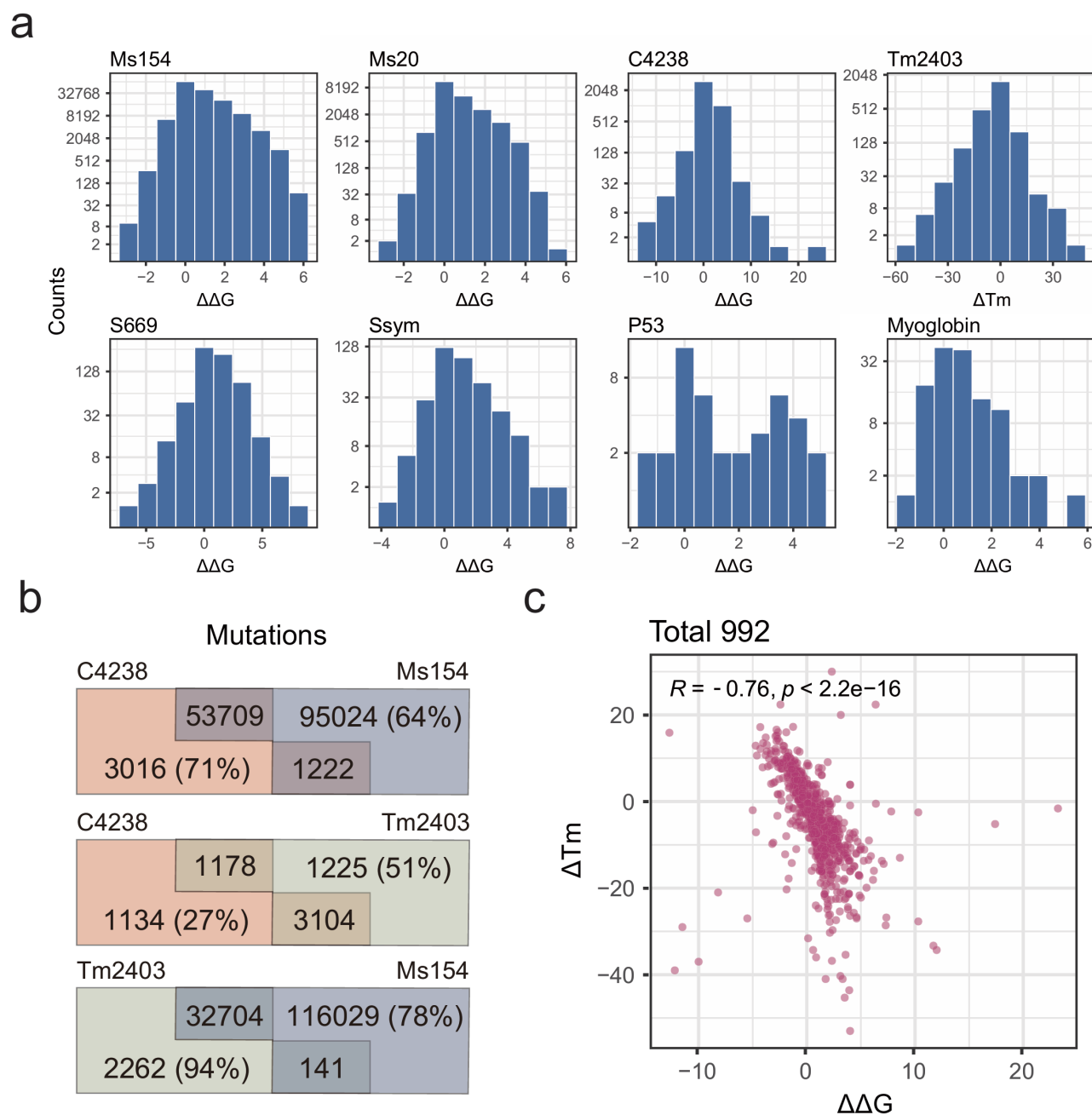
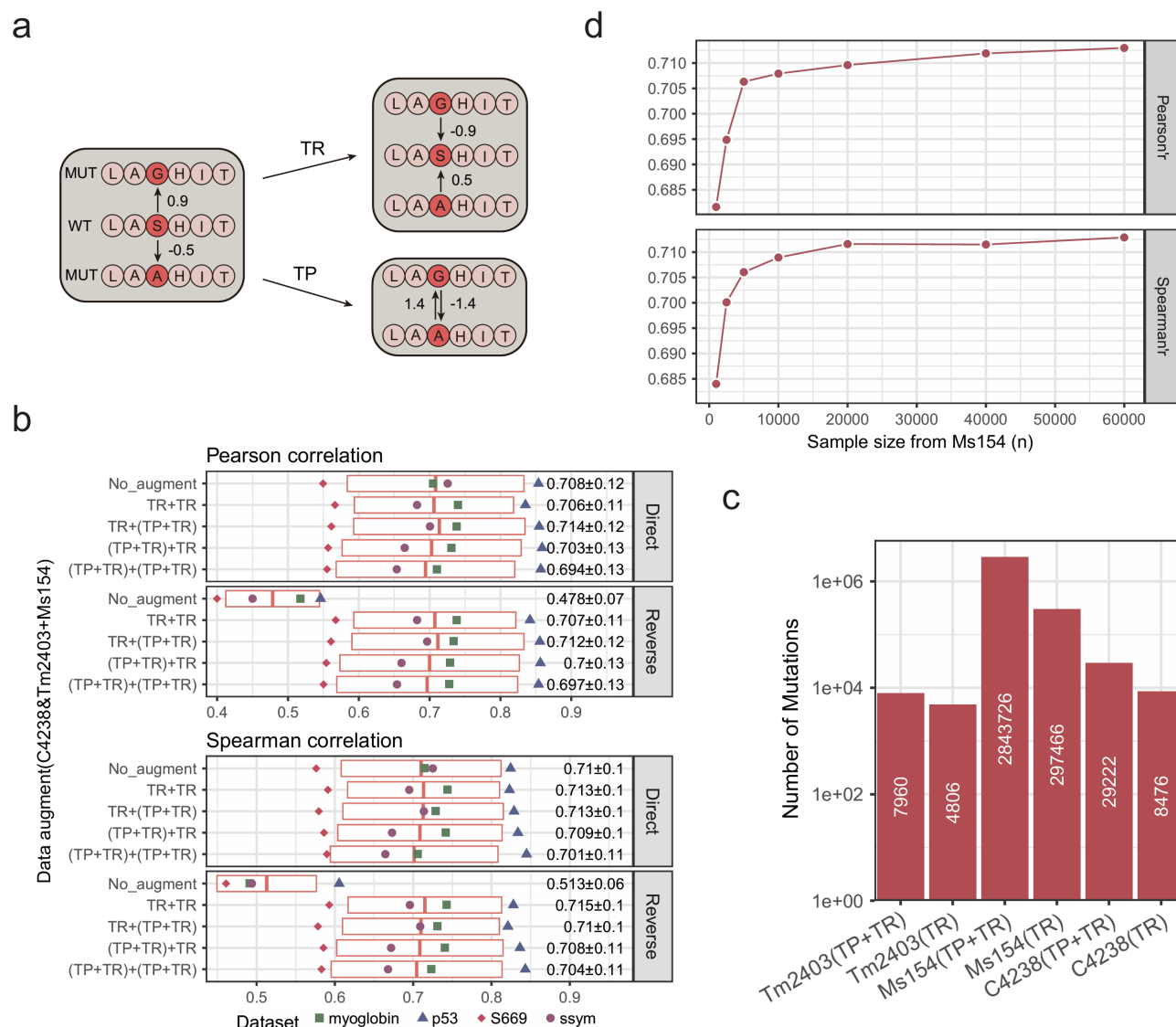


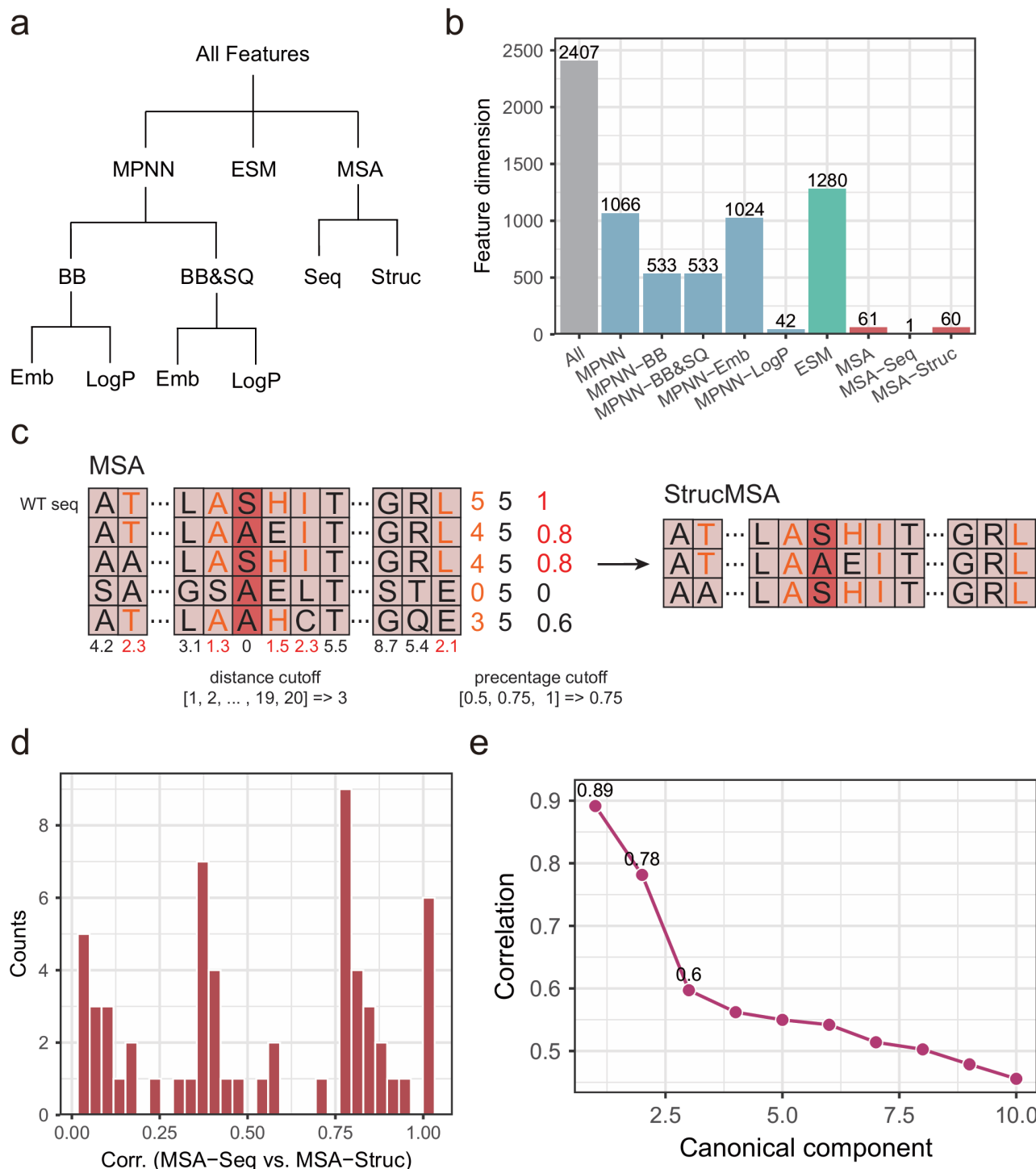
Fig. 6. Performance comparison of STAB-DKL and other methods for prioritizing stabilizing mutations. **a**, Dot plot showing the numbers of stabilizing mutations ($\Delta\Delta G < -0.5$ kcal/mol) and total numbers of experimentally characterized mutations for each of the seven proteins in the HM7 dataset. **b**, Bar chart showing mean Pearson and Spearman correlations between predicted and observed values across different proteins in the HM7 dataset for each method. Results are presented as mean $\pm$ SEM. **c**, Bar chart illustrating the performance of different modeling approaches on NDCG@5, NDCG@10, and NDCG@20 metrics. Results are presented as mean $\pm$ SEM. **d**, Bar chart illustrating the performance of different modeling approaches on Precision@1, Precision@5, and Precision@10 metrics. Results are presented as mean $\pm$ SEM. **e**, Line plots showing changes in NDCG@5, NDCG@10, and NDCG@20 under different variance offset settings.



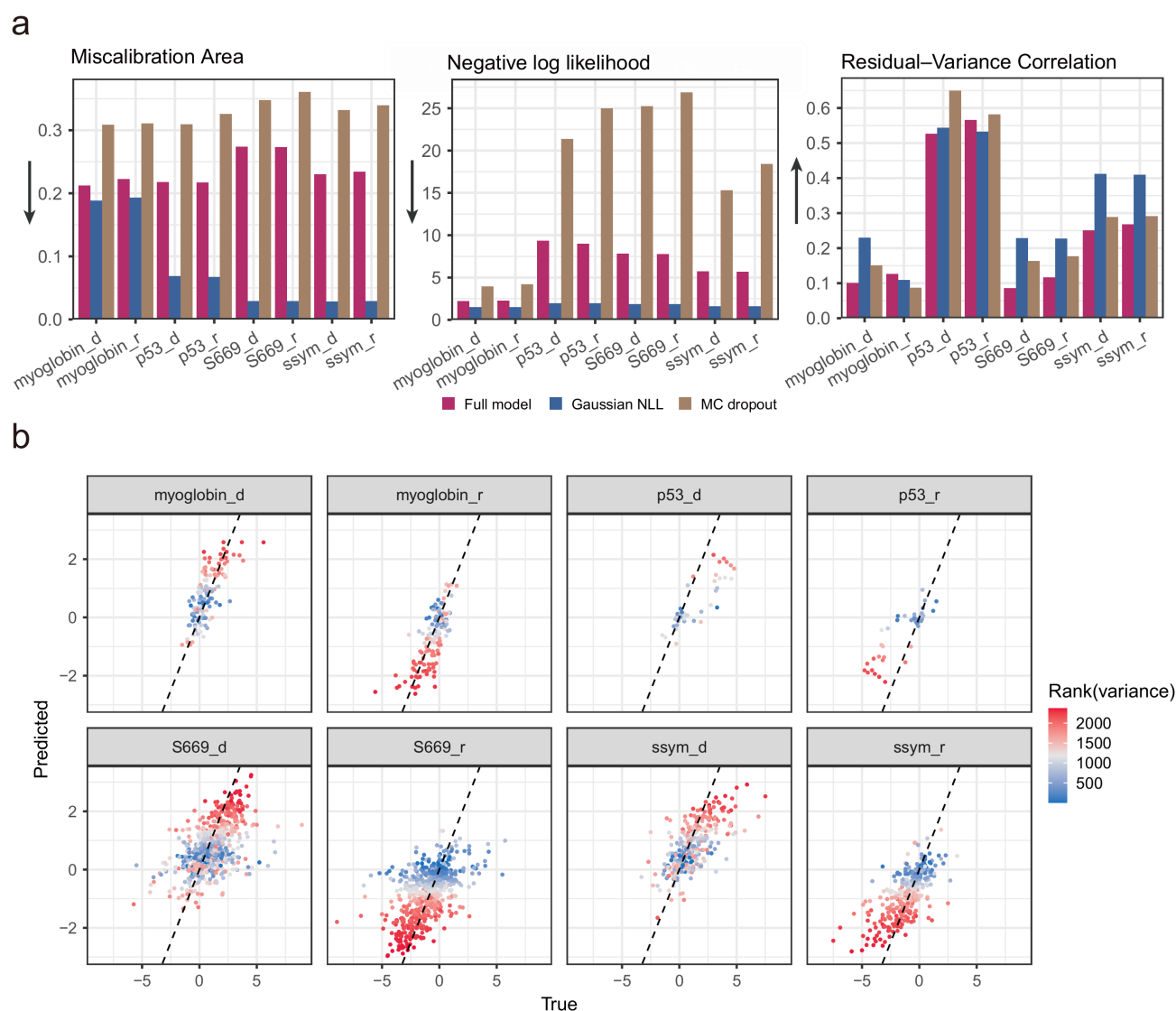
Supplementary Fig. 1. Distribution and relationship of stability-related data used for model development. **a**, Histograms showing the distribution of $\Delta\Delta G$ and ΔT_m across datasets. **b**, Mutation-level similarity among the Tm2403, C4238, and Ms154 datasets. **c**, Correlation between $\Delta\Delta G$ and ΔT_m for 992 overlapping mutations shared by the Tm2403 and C4238 datasets.



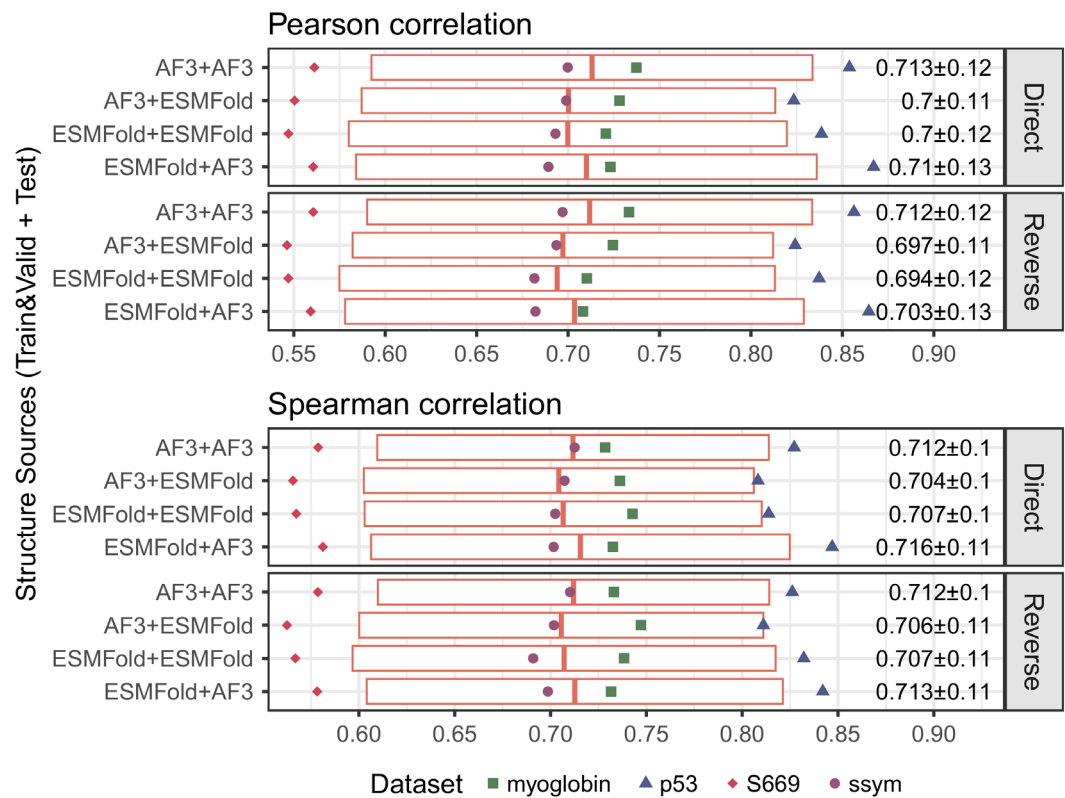
Supplementary Fig. 2. Data augmentation strategies for model development. **a**, Schematic illustration of TR and TP data augmentation strategies. **b**, Comparison of model performance under different augmentation strategies applied to the training datasets. Paired augmentation labels indicate the strategies applied to C4238/Tm2403 and Ms154, respectively. Shown as point plots with mean $\pm$ SD error bars. **c**, Bar plot illustrating dataset expansion achieved by TR and TP + TR augmentation. **d**, Line plot showing the effect of training sample size on model performance using subsets sampled from the augmented Ms154 dataset.



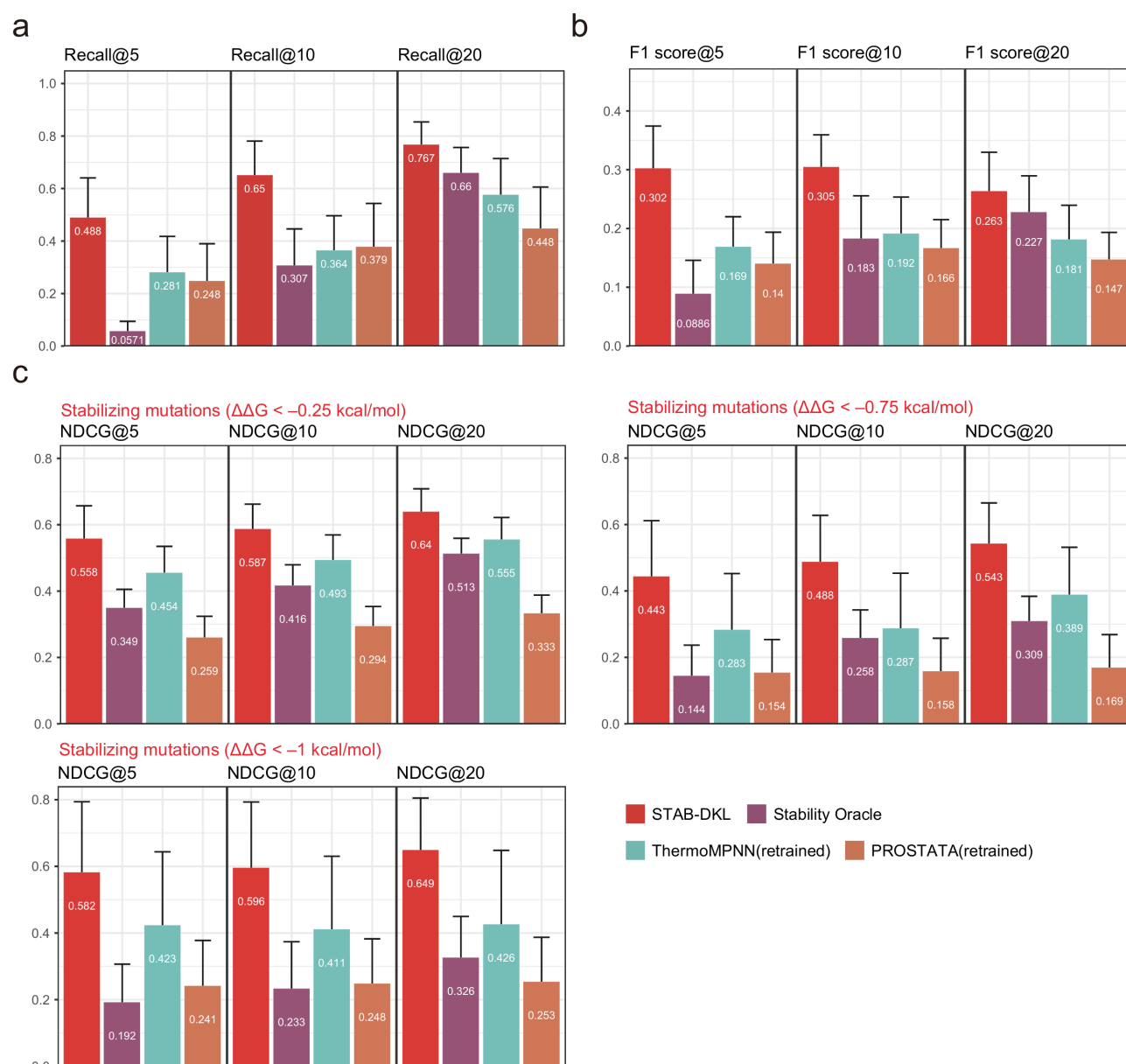
Supplementary Fig. 3. Composition and relationships of the input features. **a**, Hierarchical diagram illustrating the composition of the different input feature components. **b**, Bar plot showing the dimensionality of different feature types. **c**, Schematic overview of the StrucMSA feature extraction process. **d**, Histograms showing the distribution of Spearman correlations for MSA-Seq and MSA-Struc feature types. **e**, The first 10 canonical correlations between MSA-derived features and ESM-MSA-1b embeddings, obtained by canonical correlation analysis (CCA).



Supplementary Fig. 4. Evaluation of predictive uncertainty. **a**, Bar chart comparing uncertainty metrics including Miscalibration Area, Negative Log Likelihood, and Residual–Variance Correlation for STAB-DKL, Gaussian NLL and MC dropout. **b**, Scatter plots of STAB-DKL predictions versus observed values for different test sets, colored by the rank of predictive variance.



Supplementary Fig. 5. Effect of predicted structure sources on model performance. Comparison of model performance using different combinations of AlphaFold3- and ESMFold-predicted structures for the training/validation and test datasets. “Train&Valid” and “Test” indicate the structure sources used for training and validation, and for testing, respectively. Results are shown as point plots with mean ± SD.



Supplementary Fig. 6. Evaluation of model performance using ranking metrics. a, Bar chart illustrating the performance on Recall@5, Recall@10, and Recall@20 metrics for various methods. **b,** Bar chart illustrating the performance on F1@5, F1@10, and F1@20 metrics for various methods. **c,** Bar charts illustrating NDCG@5, NDCG@10, and NDCG@20 for different methods when using stability thresholds of -0.25 , -0.75 , and -1 . Results are presented as mean $\pm$ SEM.