Protein folding stability estimation with an explicit consideration of unfolded states

Heechan Lee[1,2] and Hahnbeom Park[1,3,*]

[1]Medical Research Division, Korea Institute of Science and Technology, Seoul, Republic of Korea
[2]School of Biological Sciences, Seoul National University, Seoul, Republic of Korea
[3]Institute for Convergence, Sungkyunkwan University, Suwon, Republic of Korea
[*]Correspondence: hahnbeom@kist.re.kr

Abstract

Thermal folding stability is a critical requirement for the vast majority of proteins. Computational methods suggested to date for the absolute folding stability (ΔG) prediction – including those driven from protein structure prediction AIs – show clear limitations on reproducing quantitative experimental values. Here we present IEFFEUM, a deep neural network that jointly estimates ΔG and the equilibrium ensemble of folded and unfolded states represented by their residue-pair distance probability distributions. This joint learning considerably enhances the ΔG prediction accuracy against the scenario where ΔG prediction was learned alone. To improve the model, we further extend the dataset compared to previous related works, which includes the Mega-scale small proteins and disordered proteins for training, as well as wild-type natural proteins with sizes up to 900 residues for thorough validation. We show that IEFFEUM is robust to various protein types and sizes, and can be applied to more general types of mutational effects such as sequence insertions or deletions.

Introduction

A majority of proteins function by folding into unique three-dimensional structures [1]. Because naturally evolved proteins are often only marginally stable in their unfolding free energy (ΔG) [2][3], a small shift in the folding stability can result in a crucial impact. For instance, structure-disrupting mutations with poorer ΔG can lead to protein misfolding and aggregation which can eventually cause human diseases [4][5] or poor protein expressions [39]. Thus, accurate ΔG quantification is highly demanding in many biotechnological applications, such as therapeutic development or biocatalysis [6].

Computational estimation of ΔG has been studied from various perspectives. One of the most popular methods these days is to approximate it using the confidence estimate from structure prediction results. For instance, pLDDT scores from AlphaFold2 [9] or ESMFold [10] have been used to filter out potentially unstable protein designs [11]. Similarly, the pseudolikelihood assigned to a sequence by protein language models (PLMs) or inverse folding models has been explored as a proxy for protein stability [12]. However, these methods are not designed to estimate ΔG, and therefore show limitations in providing a quantitative estimate. In parallel, considerable efforts have been made to build networks for the estimation of $\Delta\Delta G$, the change in stability upon mutation ($\Delta G_{MT} - \Delta G_{WT}$) [7][8]. However, despite its utility for protein engineering, it provides little clue to the absolute protein stability, and their utility is mostly restricted to single point mutations. These limitations highlight both the challenges and the need for developing models capable of accurately predicting ΔG.

We reasoned that the challenge of ΔG prediction originates from the ambiguity of the sequence-to-structure relationship at unfolded states. Because ΔG is by definition the gap between G_{folded} and $G_{unfolded}$, to accurately quantify ΔG, it is very natural to consider the unfolded state along with the folded state. In fact, sequences not only influence folded states but also unfolded states; for instance, unfolded states of poly-Leu and poly-Lys peptides cannot be equal to each other. This sequence-dependence to unfolded states was previously modeled using a linear model called "reference state energy" [40], which was quite effective yet too simple to capture the complex nature of unfolded states, or were unmodeled and left as at a "null state" regardless of protein sequences in many deep-learning approaches [12][26]. We reasoned that the key question in this endeavor should be how to explicitly abstract the unfolded state into a deep learning architecture to effectively represent the protein unfolded state.

In this work, we introduce IEFFEUM (**I**n silico **E**valuation of Un**f**olding **F**ree **E**nergy with **U**nfolded states **M**odeling), a deep learning model that addresses the challenge of accurate ΔG prediction by explicitly abstracting the unfolded state ensemble. Conceptually speaking, IEFFEUM makes two hypotheses to abstract the unfolded state representation: first, proteins fold via a two-state folding model, and second, diverse unfolded state conformation are mean-field approximated by the Flory random coil-based representation [17]. Then, it estimates ΔG from the relative thermodynamic preferences of a given sequence

at the average structures of folded and unfolded states, which corresponds to the unfolding free energy definition: ΔG = -RT ln([U]/[F]). It is demonstrated that this approximation is powerful enough to accurately predict ΔG for various types of proteins, and also for estimating $\Delta\Delta G$ from sequence deletions or insertions that previous $\Delta\Delta G$-centric methods were not capable of.

Results

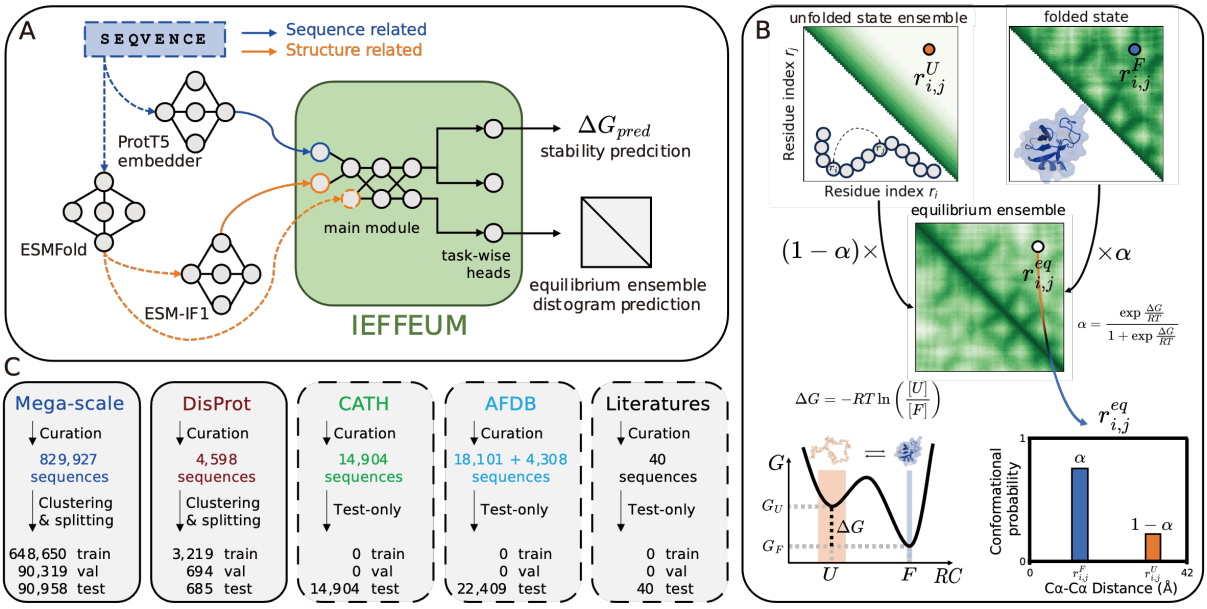

Fig. 1: Overview of IEFFEUM framework, key concepts, and datasets. (A) The IEFFEUM architecture. A schematic representation of the IEFFEUM deep learning model. Inputs include sequence embeddings from ProtT5, structure embeddings from ESM-IF1, and a pairwise Ca distogram from ESMFold. These are processed by a transformer-based main module with task-specific heads, including unfolding free energy (ΔG) and equilibrium ensemble distograms predicting heads. (B) Boltzmann distribution integration. Visual depiction of how the Boltzmann distribution governs the equilibrium ensemble. Heatmaps illustrate distance distributions for unfolded and folded states, mixed according to probabilities derived from the Boltzmann distribution using the formula for α and ΔG. The energy diagram and conformational probability bar chart further clarify the thermodynamic basis. (C) Datasets used. Overview of the Mega-scale, DisProt, CATH, AFDB, and Literature datasets employed for training and evaluation, detailing curation and splitting strategies with sequence counts for train, validation, and test sets.

Overview of IEFFEUM.

As highlighted above, the key concept of IEFFEUM is to abstract the conformations of both folded and unfolded states of proteins. We use the residue-pair Cα distance distributions (distogram) for this purpose. The sequence embeddings are brought from ProtT5 [21]. To get the folded state embedding, the structure is first predicted using ESMFold [10]. From the predicted structure ESM-IF1 [22] generates the structural embedding. This structural embedding and the pairwise Cα distogram are fed into the model as separate structural inputs (**Fig. 1A**). The predicted and experiment ΔG values are clamped into the Mega-scale dataset's experimental dynamic range of [-1, 5] kcal/mol [18]. The unfolded state Cα distogram is a simple function of sequence separations following the Flory model and hence not specifically modeled as input. Using the sequence and the folded-state structural embedding, a transformer-based module jointly updates these embeddings and the distogram

inspired by AlphaFold2's Evoformer [9]. Using the output embedding, the network jointly estimates target objectives by applying separate heads: ΔG, sequence reconstruction, and the Cα distogram. During training, the ΔG labels were brought from experimental data, and the distogram labels were assigned as a mixture of folded and unfolded state distograms, with a ratio derived from the Boltzmann distribution (**Fig. 1B**). By doing so, the network can effectively learn to what extent the given sequence can accommodate a denatured structure with respect to the folded structure ensemble. IEFFEUM was trained on a carefully collected subset of 648,650 proteins from Mega-scale dataset, which contains ΔG values inferred from protease resistance experiments with sequence lengths ranging from 30 to 80 and 3,219 disordered proteins from the DisProt database [19] with at most 70 amino acids long. Further details on the model's architecture, inputs, unfolded state ensemble modeling, and datasets are available in the Methods.

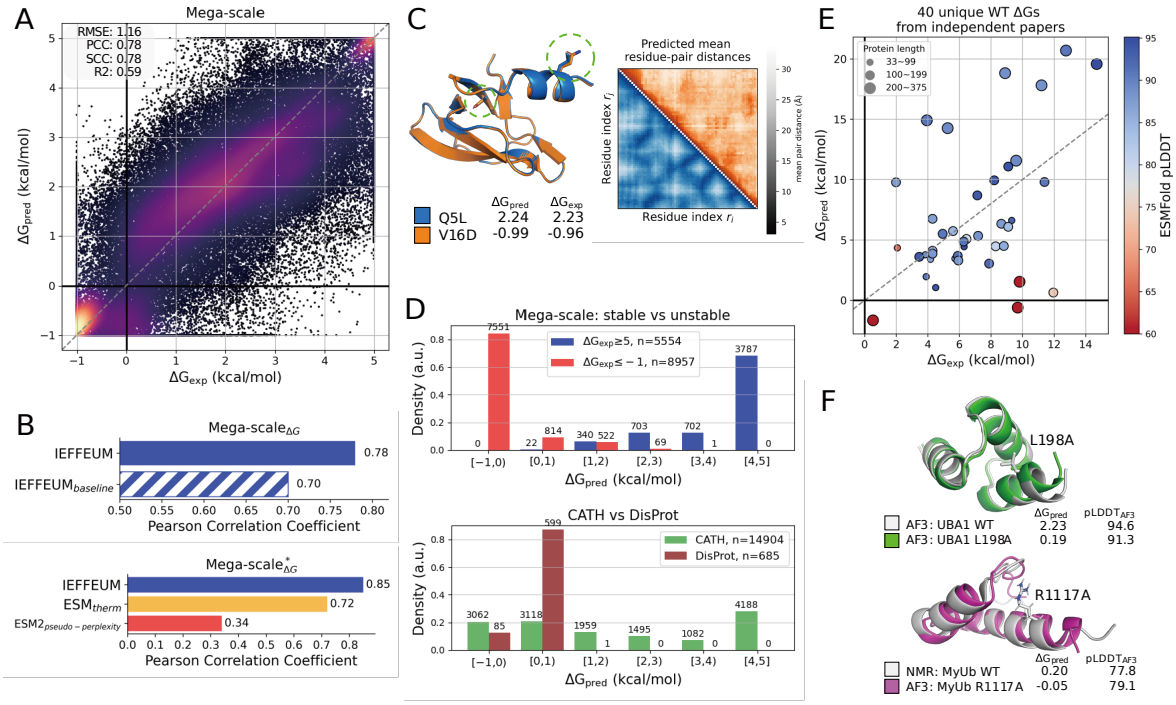

Fig. 2. IEFFEUM performance on Mega-scale and other datasets. (A) Mega-scale Dataset Prediction Accuracy. Scatter plot depicting predicted ΔG (ΔG_{pred}) versus experimentally determined ΔG (ΔG_{exp}) for proteins from the Mega-scale test set. Both values are clamped into the experimental dynamic range of [-1, 5] kcal/mol. (B) Contribution of Equilibrium Ensemble Objective. Bar chart comparing Pearson correlation coefficient (PCC) values for IEFFEUM and a baseline model (IEFFEUM without equilibrium ensemble prediction), both evaluated on the Mega-scale dataset. The enhanced PCC for IEFFEUM highlights the performance improvement gained from incorporating the equilibrium ensemble prediction objective. (C) Equilibrium Ensemble Distogram Prediction for Mutational Effects. Equilibrium ensemble distogram heatmaps for mutant proteins (Q5L, V16D; wild-type PDB ID: 3DKM) overlaid with ESMFold-predicted structures. Predicted mean residue-pair distances and corresponding ΔG values are presented for each variant. (D) Comparative Performance on Mega-scale Dataset. Violin plots illustrating the distribution of PCC values achieved by IEFFEUM, ESM_{therm}, and ESM2 when evaluated on the common Mega-scale* test set (see Methods). (E) Predicted ΔG Distributions Across Diverse Datasets. Density plots displaying the distributions of predicted ΔG values for: *Top*, stable ($\Delta G_{exp} \geq 5$ kcal/mol) versus unstable ($\Delta G_{exp} \leq -1$ kcal/mol) proteins from the Mega-scale test set. *Middle*, CATH versus DisProt datasets, representing structured domains and intrinsically disordered proteins, respectively. *Bottom*, Mammalian CATH proteins, categorized by protein length ranges to assess length-dependent trends. (F) Wild-type Protein ΔG Predictions. Scatter plot showing

predicted ΔG versus experimental ΔG for 40 unique wild-type proteins from independent literature sources. Symbol size and color gradient correspond to protein length and pLDDT scores of input structures (ESMFold predictions), respectively. (G) IEFFEUM and AlphaFold3 Performance on Disease-Related Mutations. Bar charts quantifying IEFFEUM ΔG_{pred} and AlphaFold3 pLDDT for wild-type and mutant UBA1 (L198A) and MyUb (R1117A) proteins. The comparison assesses model capabilities in capturing the destabilizing effects of mutations.

IEFFEUM can predict unfolding free energy within 1.2 kcal/mol mean error.

We first tested the accuracy of IEFFEUM predictions on a separately prepared test set from the Mega-scale dataset. These proteins have at least < 0.3 sequence identity to any of the proteins used in training. IEFFEUM achieved an RMSE of 1.2 kcal/mol and a Pearson correlation coefficient of 0.78 (**Fig. 2A***,* more detailed statistical analyses provided in **Table S1**). We think these values should approximate the current network's precision limit, at least for small proteins, at ideal conditions free from experimental disparities (e.g. pH, temperature, solvent, etc). To see how this compares to other existing tools, we ran ESM_{therm} [26], a fine-tuned version of ESM2 for Δ*G* prediction, and the ESM2 sequence pseudolikelihood (see Methods) [10]. Compared to 5,356 test cases that commonly overlap between IEFFEUM and ESM_{therm}, IEFFEUM achieved a PCC of 0.85, significantly outperforming both ESM_{therm} and ESM2 (0.72 and 0.35, respectively, **Fig. 2D**).

Training IEFFEUM with an explicit unfolded state model improves ΔG prediction.

Then, we confirmed whether this improvement resulted from the incorporation of unfolded-state modeling. Training a simpler version of IEFFEUM that does not contain any unfolded-state-related aspects ($IEFFEUM_{baseline}$), the PCC dropped from 0.78 to 0.70, with RMSE change from 1.16 to 1.39 kcal/mol. Inspecting the distogram prediction, it was highly consistent with the Δ*G* prediction; while stable proteins had ordered distograms resembling those from ESMfold, unstable ones had fuzzy distograms close to mixtures of folded and unfolded distograms (**Fig. 2C**).

We further investigated ablation studies to reveal the importance of key components of the network (**Table S1**). Removing the triangle multiplicative updates [9] or replacing all differential self-attention layers [23] with conventional self-attention significantly reduced performance. Critically, training IEFFEUM without the equilibrium ensemble prediction objective also resulted in a comparable performance decrease (**Fig. 2B**), underscoring the importance of explicitly modeling the unfolded state ensemble for accurate Δ*G* prediction.

IEFFEUM is broadly applicable to various types of proteins.

Given the promising prediction accuracy on small proteins at highly controlled conditions, we next evaluated IEFFEUM's performance on more realistic problems. First, we asked about its discriminative power in distinguishing unstable (experimental $\Delta G \leq -1$ kcal/mol) versus stable (experimental $\Delta G \geq 5$ kcal/mol) sequences within the well-controlled Mega-scale test set (**Fig. 2E***, top*). The network can discriminate between those, giving a P-value < 0.01. Next, we moved on to a more challenging but realistic problem, which was to apply the same discrimination test on semi-labeled wild type natural proteins. This dataset includes 14,904 folded sequences from CATH which contains structured domains from wild-type

proteins, and 685 disordered proteins from DisProt. We found that the distributions of predicted Δ*G* values reflected the differences in the characteristics in two distinct datasets, giving a P-value < 0.01. Investigating CATH domains with low Δ*G* (i.e. unstable) predictions, we found those are either a case of broken polypeptide, obligate oligomers, or transmembrane domains, suggesting careful usage depending on the protein context (**Fig. S3**).

We then tested its performance on labeled proteins with broader types of proteins with sizes ranging from 33 to 375 residues. A set of 40 wild-type protein sequences were collected having their experimentally determined Δ*G* values [22] (**Fig. 2F**; see Methods). Across all 40 sequences, IEFFEUM achieved a PCC of 0.47, which is poorer than that on the well-controlled data set. To understand the reason, we analyzed the impact of the quality of folded state conformation. Using the overall pLDDT from ESMFold, a clear trend was observed; restricting the analysis to the 34 sequences with pLDDT > 80 improved the PCC to 0.61; with pLDDT > 90 (16 sequences), PCC further increased to 0.73. These findings suggest that IEFFEUM's Δ*G* prediction accuracy is correlated with the confidence and quality of the input target folded state, with higher accuracy observed for proteins with well-defined structures. Further analysis between Δ*G* prediction and folded state quality is in Discussion.

IEFFEUM can capture destructive mutations AlphaFold fails.

Next, we compared IEFFEUM with AlphaFold3 server (AF3) [28] on two disease-related, structure-disrupting missense mutations: L198A in the ubiquitin-associated domain (UBA) of human Rad23 protein (hHR23a) [4] and R1117A in the Myosin VI MyUb domain [5] (**Fig. 2G**). AF3 correctly predicted the effect of L198A in UBA1 (**Fig. 2G***, top*) but the absolute value does not well match the experimental Δ*G* (before and after both are > 90), and it failed to predict the effect of R1117A in MyUb (**Fig. 2G***, bottom*). Notably, AF3 predicted structures for both mutants very similar to their wild-type structures despite the known structure-disrupting nature of these mutations. In contrast, IEFFEUM not only correctly predicted the negative effect of both mutations but also predicted Δ*G* values near 0 kcal/mol, suggesting significant destabilization of the folded state. These examples highlight the limitation of using structure prediction methods for absolute stability estimation, and also the need for a separate Δ*G* predictor.

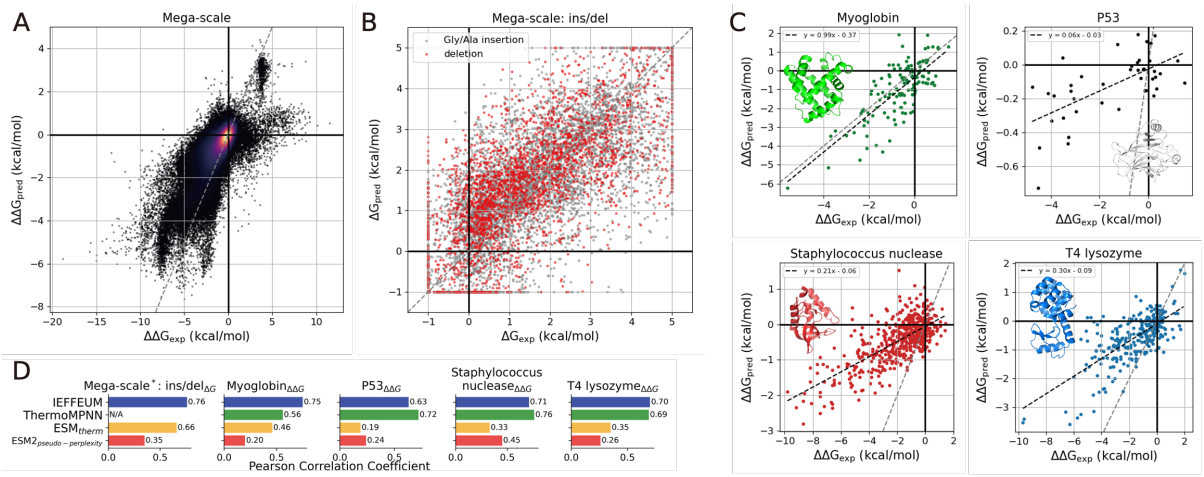

Fig. 3: Comparison of IEFFEUM with Other ΔΔ*G* Predictors. (A) Mega-scale (full). Scatter plot of predicted ΔΔ*G* versus experimental ΔΔ*G* for substitution and insertion/deletion mutations on the Mega-scale dataset. (B) Mega-scale (insertions/deletions). Scatter plot of predicted Δ*G* versus experimental Δ*G* for insertion and deletion mutations on the Mega-scale dataset. (C) Case Study ΔΔ*G* Distributions. Scatter plots comparing predicted ΔΔ*G* distributions for IEFFEUM, ThermoMPNN, ESMtherm, and ESM2 for Myoglobin, P53, Staphylococcal Nuclease, and T4 Lysozyme case studies. (D) PCC Comparison Across Methods. Bar chart summarizing Pearson Correlation Coefficients (PCC) for IEFFEUM, ThermoMPNN, ESMtherm, and ESM2 on insertion/deletion mutations and case study datasets.

IEFFEUM can accurately predict ΔΔG for substitutions, insertions, and deletions

IEFFEUM exhibited strong predictive performance for ΔΔ*G*, encompassing both substitutions and insertions/deletions (**Fig. 3A** and **3B**). To further benchmark this capability, we compared IEFFEUM against ThermoMPNN, ESMtherm, and ESM2 (**Fig. 3D**). For insertions and deletions, a direct comparison with ThermoMPNN was not applicable. However, IEFFEUM demonstrated a clear advantage (PCC=0.76) over both ESMtherm (PCC=0.66) and ESM2 (PCC=0.35) predicting Δ*G* on a common test set of insertion and deletion mutations. This performance suggests that IEFFEUM is better suited to capturing the more extensive structural perturbations often caused by insertions and deletions compared to substitutions.

Next, we evaluated the models on four case studies: myoglobin [29], p53 [30], staphylococcal nuclease [31], and T4 lysozyme [31]. IEFFEUM achieved PCCs of 0.75, 0.63, 0.71, and 0.70, respectively (**Fig. 3C**), comparable to ThermoMPNN (0.56, 0.72, 0.76, and 0.69, respectively). ESM_{therm} and ESM2 showed relatively poor performance in all four case studies, with specific PCC values reported in **Fig. 3D.**

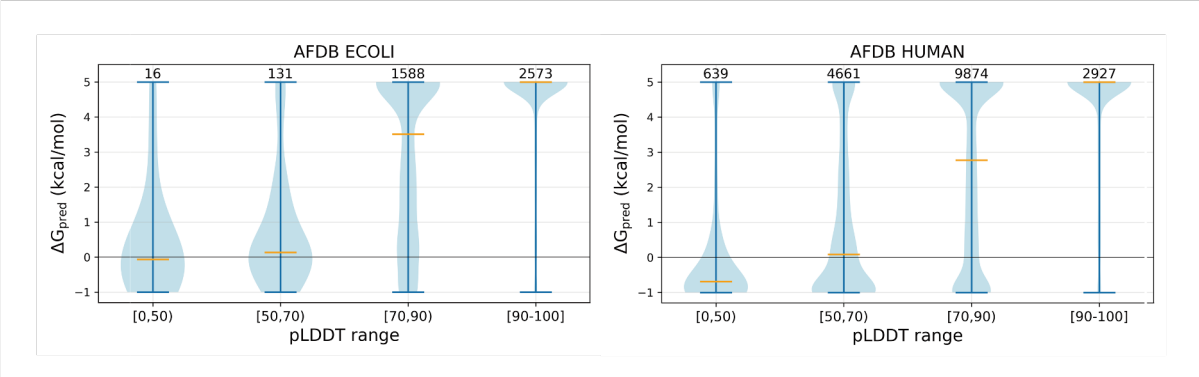

Fig. 4: IEFFEUM performance dependency on input structure quality. Violin plots showing distributions of predicted ΔG values for E. coli and human proteome sequences from AFDB, categorized by pLDDT ranges (very low, low, confidence and very high, respectively). Demonstrates the correlation between input structure quality (pLDDT) and predicted protein stability.

Evaluation of IEFFEUM on AFDB

To evaluate IEFFEUM's performance dependency on input structure quality, we predicted Δ*G* values for E. coli and human proteome sequences using AlphaFold2-predicted structures from the AFDB (**Fig. 4**). A clear correlation emerged between the structures' pLDDT scores and the predicted ΔG values. Lower pLDDT structures were associated with less stable ΔG predictions, while higher pLDDT structures correspond to more stable predictions.

Discussion

The advantage of explicitly considering unfolded state ensemble.

Our results suggest that explicitly considering the unfolded state ensemble contributes to improved accuracy in Δ*G* prediction. The observed performance decrement when neglecting the equilibrium ensemble objective was comparable to that seen with ablations of key architectural components. This highlights the relevance of modeling the unfolded state ensemble for achieving accurate Δ*G* predictions, aligning with the thermodynamic basis of Δ*G* as the free energy difference between folded and unfolded states, an advantage over many ΔΔ*G* predictors whose reliance on a single folded state (typically the wild-type structure) may limit their ability to capture the full thermodynamic impact of mutations [7][8].

Transfer learning of large pre-trained models can improve model robustness.

IEFFEUM predicts Δ*G* and the equilibrium ensemble by integrating sequence and structure embeddings from ProtT5 and ESM-IF1, along with structural input. Leveraging transfer learning from extensive datasets, including 2.1 billion and 45 million protein sequences in BFD and UniRef50 (used to train ProtT5), respectively, and 12 million AlphaFold2-predicted structures (used to train ESM-IF1), likely contributes to mitigating overfitting in IEFFEUM, a strategy also seen in ThermoMPNN [7]. This pre-training approach may offer an advantage over methods like $\text{ESM}_{\text{therm}}$, which, in contrast, showed significant overfitting when fine-tuning the entire ESM2 model on a single dataset [26].

IEFFEUM is affected by the folded state structure quality.
IEFFEUM's performance on the various test sets reflects its ability to capture protein characteristics. Beyond mutational datasets, IEFFEUM also showed an ability to differentiate ΔG values for wild-type proteins, achieving a PCC of 0.73 when provided with higher quality folded structures (pLDDT > 90) for proteins up to 375 residues. This dependence on input structure quality, while seemingly a factor, can be interpreted as consistent with the hypothesis that accurate ΔG prediction necessitates considering the full thermodynamic landscape, including both folded and unfolded states. Further analysis reinforced the role of the folded state in capturing the full folding landscape (**Fig. S4**). Substituting the structural inputs for low-pLDDT ESMFold structures with high confidence AF3 structures. These AF3 structures, closely resembling corresponding crystal structures, served as improved representations of the "target folded state." When IEFFEUM was provided with these enhanced structural inputs, the predicted ΔG values increased (more stable). This observation further supports the interpretation that IEFFEUM effectively integrates information from both folded and unfolded state ensembles for ΔG estimation.

Interpretable predictions and applications.
Given IEFFEUM's per-residue ΔG contributions, we explored its potential for guiding de novo protein design. Preliminary results suggest that focusing redesign on predicted destabilizing residues can be an effective strategy, potentially leading to improved protein stability.

Limitations of IEFFEUM.
While IEFFEUM demonstrates promising performance in protein stability prediction, several inherent limitations warrant consideration. First, the primary training dataset, Mega-scale, possesses an experimental dynamic range of [-1, 5] kcal/mol, with protein sizes ranging up to 80 residues [18]. This restricted range, inherent to the proteolysis-based experimental methodology, may constrain IEFFEUM's capacity for accurate extrapolation beyond these boundaries, potentially limiting its performance for proteins exhibiting exceptionally high or low stability. Second, IEFFEUM's framework is built on a simplified two-state folding model. While this approximation is widely used [18], it may not fully capture the complexities of the folding landscapes of all proteins, particularly larger or multi-domain proteins, potentially affecting prediction accuracy in such cases. Furthermore, the equilibrium ensemble representation employed for training, while computationally expedient, represents a simplified, discrete approximation of what is likely a more continuous conformational distribution. This simplification could limit the model's ability to resolve fine-grained details of protein conformational ensembles. Third, the network does not apply to membrane proteins and obligate oligomers because it was mainly trained on proteins that are water-soluble and monomeric. We intend to address this in further studies, specifically to predict proteins in their natural conditions. Finally, as IEFFEUM relies on transfer learning from large-scale pre-training, its performance is intrinsically linked to the quality and scope of these foundational models and datasets. While transfer learning likely mitigates overfitting, the model's standalone utility without this pre-training remains to be fully established,

highlighting a dependence on external data resources. These limitations, while acknowledged, also define key directions for future model refinement and development aimed at enhancing the scope and accuracy of protein stability predictions.

Conclusion and Future Directions

In summary, IEFFEUM offers a method for ΔG and equilibrium ensemble prediction, incorporating the unfolded state ensemble and leveraging transfer learning. While these findings are encouraging, further research is needed to fully explore the model's capabilities and limitations, and to guide future development in protein stability prediction and design.

Methods

IEFFEUM Architecture. IEFFEUM is a transformer-based model designed to predict multiple protein properties, including the unfolding free energy (ΔG), the equilibrium ensemble of conformational states, and sequence recovery. The model takes three inputs: a sequence embedding of shape [N, 1024] generated by ProtT5 embedder, a one-hot encoded, binned pairwise distance map (distogram) of shape [N, N, 21] representing the folded structure (predicted by ESMFold unless otherwise stated), and a structure embedding of shape [N, 512] extracted from ESM-IF1, where N is the sequence length of target protein. The IEFFEUM architecture comprises three main components: input embeddings attention layers, a main transformer module incorporating triangle multiplicative updates, and task-specific simple output multilayer perceptron (MLP).

Initially, IEFFEUM reshapes the sequence embedding and the structure embedding respectively to a common dimension of [N, 21] using differential self-attention. These reshaped embeddings are then concatenated into a tensor of shape [2, N, 21].

The main module is a modified version of AlphaFold2's Evoformer, where triangle attention is removed due to its heaviness and all self-attention layers are replaced with differential self-attention. During the main blocks, the pairwise distogram and the concatenated embeddings are jointly updated through differential self-attention with pair bias and outer product mean as similar to the original Evoformer. IEFFEUM's main module comprises 11 of these main blocks.

Finally, an MLP with two hidden layers predicts the final output for each objective. Specifically, ΔG and sequence recovery are predicted from the main module-updated embeddings, while the equilibrium ensemble is predicted from the main module-updated pairwise distogram. IEFFEUM predicts a protein's ΔG by predicting per-residue ΔG contributions and then through an unweighted summation of these contributions. We observed that optimizing destabilizing sequences of a protein based on per-residue contribution results in a higher predicted ΔG than naively redesigning the whole sequence, suggesting potential usage of IEFFEUM on de novo protein design (**Fig. S1**). IEFFEUM can also predict $\Delta\Delta G$ by subtracting the predicted wild-type ΔG from a predicted mutant ΔG, obtained from independent runs of the model.

Equilibrium Ensemble Modeling. The Flory random coil model (Equation 1) is a well-established model for describing the unfolded state ensemble of proteins. This model provides a statistical description of the average distances expected between residues in a flexible, disordered chain. Therefore, we employed the Flory random coil model with fitted parameters [35] [36] (Equation 2) to calculate the root-mean-square (RMS) pairwise Cα distances between residue indices *i* and *j* in the unfolded states ($r_{i,j}^{U}$) of the protein. We use these RMS pairwise distances as an explicit representation of the spatial extent of unfolded states to model their contribution to the conformational ensemble.

$$D = R_0 N^{\upsilon} \quad [1]$$

$$\sqrt{\langle\langle {r_{i,j}^{U}}^2 \rangle\rangle} = \sqrt{6} * 1.927 * |i - j|^{0.598} \quad [2]$$

Protein structure prediction AIs often exhibit insensitivity to structure-disrupting mutations, sometimes predicting structures for mutated sequences that are nearly identical to the wild-type structure. Therefore, we used the ESMFold-predicted structure as a proxy for the target folded (near-native) state, recognizing the limitations of AI models in capturing mutational effects. Pairwise Cα distances between residue indices *i* and *j* can be directly calculated from this modeled structure.

Assuming a simple two-state equilibrium between folded and unfolded states, we constructed our model by incorporating the Flory random coil model for unfolded conformations and the ESMFold-predicted structure as a proxy for folded conformations. Within this model, the distance between any two residues was represented as a weighted average of their folded and unfolded states distances. This weighting factor, reflecting the equilibrium ratio between folded and unfolded states (Equation 4), can be derived directly from the protein's unfolding free energy, ΔG (Equation 3). Consequently, using this equilibrium ratio we calculated the distribution of pairwise distances between residues *i* and *j* in folded and unfolded states (Equation 5), which is then represented as a distogram with a range of [2, 42] and a bin width of 2 Å (**Fig. 1B**). IEFFEUM was trained to predict this equilibrium ensemble distogram.

$$\Delta G = -RT\, ln\left(\frac{[U]}{[F]}\right) \quad [3]$$

$$[F] : [U] = exp\left(\frac{\Delta G}{RT}\right) : 1 \quad [4]$$

$$p(r_{i,j}^{F}) = \frac{exp\left(\frac{\Delta G}{RT}\right)}{1 + exp\left(\frac{\Delta G}{RT}\right)}, \qquad p(r_{i,j}^{U}) = \frac{1}{1 + exp\left(\frac{\Delta G}{RT}\right)} \quad [5]$$

Training Configuration. IEFFEUM was trained on a single A6000 48 GB GPU using an AdamW with a weight decay of 0.01 and a learning rate of 1 x 10^{-4}. The loss function was a weighted combination of three components (Equation 6, L): a Gaussian negative log-likelihood loss for ΔG prediction ($L_{\Delta G}$), and cross-entropy loss for equilibrium ensemble modeling ($L_{ensemble}$), a cross-entropy loss for sequence reconstruction (L_{seq}). During training, we used a batch size of 80, increasing it to 100 for validation and testing, and reducing it to 1 for test-only datasets. Data was shuffled at the start of each training epoch, and masks were carefully employed during calculation to enable batch processing of sequences with varying lengths. We acknowledge that we did not explicitly evaluate the impact of different mask sizes on prediction performance.

$$L = L_{\Delta G} + 100\, L_{ensemble} + L_{seq} \qquad [6]$$

Dataset Preparation. The Mega-scale dataset was first curated by removing data with dG_ML values marked as unreliable (dG_ML = "—"). We then filtered out scrambled sequences with dG_ML values greater than 0.5 kcal/mol to avoid potential artifacts caused by aggregation. This curation process resulted in a Mega-scale dataset containing 829,927 sequences, with dG_ML values used as the training target.

The DisProt dataset was curated by collecting sequences shorter than 70 amino acids to enrich for IDPs. Longer sequences in the raw dataset often represented stable proteins with flexible loops, not primarily IDPs. This curation yielded a DisProt dataset with 4,705 sequences and this size of curated subset is noteworthy as it is comparable to, or even larger than, training sets commonly employed in similar studies. Although the DisProt dataset lacks experimental ΔG labels, it is a database of proteins known to be disordered or unstable. Therefore, during the training of IEFFEUM, we incorporated the DisProt dataset by modifying the loss function to encourage the model to predict ΔG values lower than 0.5 kcal/mol for these sequences. This approach implicitly guides the model to learn features associated with disordered proteins, even in the absence of explicit ΔG values for DisProt. We reasoned that by training IEFFEUM to predict low ΔG values for sequences known to be disordered, the model would better capture the characteristics of unfolded or unstable protein states.

The CATH database was curated by first removing sequences with 'X', then collecting sequences shorter than 900 amino acids due to computational limitations when using ESMFold for structure prediction (specifically, CUDA out-of-memory issues encountered with an A6000 GPU with 48GB of memory). This curation resulted in a CATH dataset with 14,904 sequences for testing purposes.

The AFDB was curated by collecting predicted structures for *E. coli* and human sequences shorter than 1,000 amino acids. This yielded an AFDB dataset with 22,409 sequences and corresponding structures for testing purposes.

The S669 dataset was curated by collecting sequences and their corresponding experimental ΔG values of wild-type proteins from the original 92 papers. The dataset was first filtered to include only entries with experimental pH values between 7.0 and 7.5. Subsequently, we removed homooligomerizing sequences based on UniProt annotations [37], as well as any sequences with greater than 0.3 sequence identity to those in the Mega-scale test set. This yielded a curated, test-only dataset containing 40 sequences and experimental ΔG values.

We also used the same four case studies, p53, myoglobin, staphylococcal nuclease, and T4 lysozyme, as previously employed by ThermoMPNN as test-only datasets.

Dataset Splitting. We used the MMseqs2 easy-cluster [38] to cluster the Mega-scale and DisProt datasets with a minimum sequence identity cutoff of 0.3 The clusters were then split into train and validation/test sets at an approximate ratio of 8:2. The validation and test sets were then split randomly in half. The exact values can be found in **Fig. 1C**.

A set of 5,356 sequences, common to both the IEFFEUM Mega-scale test set and the ESM_{therm} test-set-only domains, was assembled and used for IEFFEUM and ESM_{therm} comparison (**Fig. 2D** and **Fig. 3D**; noted as Mega-scale*).

Ablation Results. All modifications were trained on the same training sets with the same hyperparameters. The best model for each modification was selected based on the highest coefficient of determination on the validation set. All modifications were evaluated with the same test sets to calculate the performance metrics.

Other Predictors. ThermoMPNN is a graph neural network, a transfer-learned version of ProteinMPNN, trained on the Mega-scale dataset to predict $\Delta\Delta G$ from point mutations. It is important to note that ThermoMPNN is designed for point mutations only and does not predict the effects of insertions or deletions, as it requires the wild-type and mutant protein sequences to be of the same length. We generated $\Delta\Delta G$ values for comparison using the Google Colab implementation of ThermoMPNN.

ESM2 is a protein language model trained on the UniRef database to predict masked amino acids in protein sequences. It has demonstrated the ability to predict the functional effects of mutations without requiring any specific training (zero-shot prediction). To quantify this predictive power on predicting ΔG, we calculated the ESM2 pseudolikelihood for a sequence of length N by summing the log probabilities of predicting each residue r_i when it is masked (Equation 7). We used esm2_t36_3B_UR50D to generate these pseudolikelihood values for our comparisons.

$$ESM2\ pseudolikelihood = \sum_{i=1}^{N} \ln P_{ESM2}\,(r_i) \qquad [7]$$

ESM_{therm} is a protein language model, a fine-tuned version of ESM2, trained on the Mega-scale dataset to predict ΔG of proteins. We used the original script provided in the ESM_{therm} GitHub repository to generate ΔG values for our comparisons.

Data, Materials, and Software Availability

The code for IEFFEUM will be publicly available on GitHub (https://github.com/HParklab/IEFFEUM). All data used in this study can be found in the original research publications: Mega-scale [18], DisProt [19], CATH [20], AFDB [21], S669 [22], UBA1 [4], MyUb [5], myoglobin [29], p53 [30], staphylococcus nuclease [31], and T4 lysozyme [31].

References

1. Shoichet, B.K., et al., *A relationship between protein stability and protein function.* Proceedings of the National Academy of Sciences, 1995. **92**(2): p. 452-456.
2. Dill, K.A., *Dominant forces in protein folding.* Biochemistry, 1990. **29**(31): p. 7133-7155.
3. Taverna, D.M. and R.A. Goldstein, *Why are proteins marginally stable?* Proteins: Structure, Function, and Bioinformatics, 2002. **46**(1): p. 105-109.
4. Wang, Q., et al., *Ubiquitin recognition by the DNA repair protein hHR23a.* Biochemistry, 2003. **42**(46): p. 13529-13535.
5. He, F., et al., *Myosin VI contains a compact structural motif that binds to ubiquitin chains.* Cell reports, 2016. **14**(11): p. 2683-2694.
6. Thomson, R.E., S.E. Carrera-Pacheco, and E.M. Gillam, *Engineering functional thermostable proteins using ancestral sequence reconstruction.* Journal of Biological Chemistry, 2022. **298**(10).
7. Dieckhaus, H., et al., *Transfer learning to leverage larger datasets for improved prediction of protein stability changes.* Proceedings of the National Academy of Sciences, 2024. **121**(6): p. e2314853121.
8. Diaz, D.J., et al., *Stability Oracle: a structure-based graph-transformer framework for identifying stabilizing mutations.* Nature Communications, 2024. **15**(1): p. 6170.
9. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* nature, 2021. **596**(7873): p. 583-589.
10. Lin, Z., et al., *Evolutionary-scale prediction of atomic-level protein structure with a language model.* Science, 2023. **379**(6637): p. 1123-1130.
11. Bennett, N.R., et al., *Improving de novo protein binder design with deep learning.* Nature Communications, 2023. **14**(1): p. 2625.
12. Cagiada, M., S. Ovchinnikov, and K. Lindorff-Larsen, *Predicting absolute protein folding stability using generative models.* Protein Science, 2025. **34**(1): p. e5233.
13. Li, F.-Z., et al., *Feature reuse and scaling: Understanding transfer learning with protein language models.* bioRxiv, 2024: p. 2024.02. 05.578959.

14. Elnaggar, A., et al., *Prottrans: Toward understanding the language of life through self-supervised learning.* IEEE transactions on pattern analysis and machine intelligence, 2021. **44**(10): p. 7112-7127.
15. Hsu, C., et al. *Learning inverse folding from millions of predicted structures*. in *International conference on machine learning*. 2022. PMLR.
16. Wales, D.J. and T.V. Bogdan, *Potential energy and free energy landscapes*. 2006, ACS Publications. p. 20765-20776.
17. Alston, J.J., et al., *The analytical Flory random coil is a simple-to-use reference model for unfolded and disordered proteins.* The Journal of Physical Chemistry B, 2023. **127**(21): p. 4746-4760.
18. Tsuboyama, K., et al., *Mega-scale experimental analysis of protein folding stability in biology and design.* Nature, 2023. **620**(7973): p. 434-444.
19. Aspromonte, M.C., et al., *DisProt in 2024: improving function annotation of intrinsically disordered proteins.* Nucleic Acids Research, 2024. **52**(D1): p. D434-D441.
20. Dawson, N.L., et al., *CATH: an expanded resource to predict protein function through structure and sequence.* Nucleic acids research, 2017. **45**(D1): p. D289-D295.
21. Varadi, M., et al., *AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences.* Nucleic acids research, 2024. **52**(D1): p. D368-D375.
22. Pancotti, C., et al., *Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset.* Briefings in Bioinformatics, 2022. **23**(2): p. bbab555.
23. Ye, T., et al., *Differential transformer.* arXiv preprint arXiv:2410.05258, 2024.
24. Buel, G.R. and K.J. Walters, *Can AlphaFold2 predict the impact of missense mutations on structure?* Nature structural & molecular biology, 2022. **29**(1): p. 1-2.
25. Shortle, D., *Propensities, probabilities, and the Boltzmann hypothesis.* Protein science, 2003. **12**(6): p. 1298-1302.
26. Chu, S.K., K. Narang, and J.B. Siegel, *Protein stability prediction by fine-tuning a protein language model on a mega-scale dataset.* PLOS Computational Biology, 2024. **20**(7): p. e1012248.
27. Meier, J., et al., *Language models enable zero-shot prediction of the effects of mutations on protein function. BioRxiv, 2021.07. 09.450648*. 2021.
28. Abramson, J., et al., *Accurate structure prediction of biomolecular interactions with AlphaFold 3.* Nature, 2024: p. 1-3.
29. Kepp, K.P., *Towards a "Golden Standard" for computing globin stability: Stability and structure sensitivity of myoglobin mutants.* Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 2015. **1854**(10): p. 1239-1248.
30. Pires, D.E., D.B. Ascher, and T.L. Blundell, *mCSM: predicting the effects of mutations in proteins using graph-based signatures.* Bioinformatics, 2014. **30**(3): p. 335-342.
31. Musil, M., et al., *FireProt 2.0: web-based platform for the fully automated design of thermostable proteins.* Briefings in Bioinformatics, 2024. **25**(1): p. bbad425.

32. Cao, L., et al., *Design of protein-binding proteins from the target structure alone.* Nature, 2022. **605**(7910): p. 551-560.
33. Dauparas, J., et al., *Robust deep learning–based protein sequence design using ProteinMPNN.* Science, 2022. **378**(6615): p. 49-56.
34. Sumida, K.H., et al., *Improving protein expression, stability, and function with ProteinMPNN.* Journal of the American Chemical Society, 2024. **146**(3): p. 2054-2061.
35. Kohn, J.E., et al., *Random-coil behavior and the dimensions of chemically unfolded proteins.* Proceedings of the National Academy of Sciences, 2004. **101**(34): p. 12491-12496.
36. De Gennes, P., *Scaling Concepts in Polymer Physics*. 1979: Cornell University Press.
37. *UniProt: the universal protein knowledgebase.* Nucleic acids research, 2017. **45**(D1): p. D158-D169.
38. Steinegger, M. and J. Söding, *MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.* Nature biotechnology, 2017. **35**(11): p. 1026-1028.
39. Hargrove, J.L. and F.H. Schmidt, *The role of mRNA and protein stability in gene expression.* The FASEB journal, 1989. **3**(12): p. 2360-2370.
40. Alford, R.F., et al., The Rosetta all-atom energy function for macromolecular modeling and design. Journal of chemical theory and computation, 2017. 13(6): p. 3031-3048.