Biophysically Grounded Deep Learning Improves Protein–Protein $\Delta\Delta G$ Prediction

Jonathan Feldman[*,a, b], **Antoine Maechler**[*,a, c], **Dianzhuo Wang**[*,a], **and Eugene Shakhnovich**[a]

[a]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA; [b]College of Computing, Georgia Institute of Technology; Atlanta, GA, USA; [c]School of Engineering, Stanford University; Stanford, CA, USA; [*]These authors contributed equally to this work.

Predicting how mutations alter protein–protein binding affinity ($\Delta\Delta G$) is central to protein engineering and therapeutic design, yet current deep learning models often fail to generalize beyond narrow benchmarks. We show that SKEMPI2, the field's dominant training and evaluation dataset, suffers from widespread and previously underappreciated data leakage caused by deep sequence and structural redundancy, leading to inflated performance estimates. We introduce ProtBFF (Protein Biophysical Feature Framework), an encoder-agnostic module that injects five interpretable biophysical priors into residue-level representations via cross-embedding attention. ProtBFF consistently improves performance under homology-based-clustering evaluation and enables general-purpose encoders, such as ESM, to match or surpass state-of-the-art specialized models or larger size models. These results demonstrate that incorporating biophysical priors could be key to robust protein property prediction.

protein–protein interactions | protein design | binding energy | deep learning | biophysical modeling

Correspondence: ***johnwang@g.harvard.edu***
shakhnovich@chemistry.harvard.edu

Introduction

Predicting how multiple mutations alter protein–protein binding affinity ($\Delta\Delta G$) is a central challenge in computational biology with direct implications for protein engineering andSanavia et al. (2020); Pucci et al. (2022); Huot et al. (2025); Wang et al. (2024). Biophysics-based methods such as molecular dynamics have long been used for this task, but they are computationally expensive and rely on hard-coded mathematical models, which limits their scalabilitySchymkowitz et al. (2005); Park et al. (2016); Alford et al. (2017). While deep learning has transformed other realms of protein science, such as protein structure prediction Abramson et al. (2024); Hayes et al. (2024); Watson et al. (2023); Feldman and Skolnick (2025), the prediction of $\Delta\Delta G$ remains constrained by small, biased datasets and by models that often fail to generalize to previously unseen proteins Hummer et al. (2025); Loux et al. (2024). This shortcoming arises from two main factors. First, the quality and quantity of available experimental measurements are limited, and existing datasets cover only a narrow diversity of proteins. Second, many models tend to overfit dataset-specific patterns rather than learning the underlying biophysical principles that govern protein–protein interactions.

A critical example is the SKEMPI2 dataset, the field's most used benchmark. While SKEMPI2 contains roughly 350 curated protein complexes, it suffers from substantial sequential and structural redundancy, with many complexes exhibiting high structural and sequence similarity Bushuiev et al. (2024); Jankauskaitė et al. (2018). As this work will show, such hidden redundancy introduces significant data leakage between training and test sets, inflating reported performance and obscuring the true generalization gap in current methods Bushuiev et al. (2024); Jankauskaitė et al. (2018).

We combine the mechanistic insights of biophysical modeling with the statistical power of large-scale deep learning. To this end, we introduce ProtBFF (Protein Biophysical Feature Framework), an encoder-agnostic module that enriches embedding-based predictors with explicit biophysical features. ProtBFF modulates residue-level embeddings based on their local structural and physicochemical context and integrates wild-type and mutant representations through cross-embedding attention. By injecting biophysical priors directly into the representation space, ProtBFF biases learning toward features that are informed by known physical determinants of protein–protein interactions while retaining the flexibility of data-driven models.

Unlike prior work that proposes entirely new $\Delta\Delta G$ architectures Luo et al. (2023); Yu et al. (2024); Mo et al. (2024) or relies solely on physics-based scoring functions Schymkowitz et al. (2005); Alford et al. (2017), ProtBFF acts as a plug-in that integrates seamlessly with any pretrained encoder. We demonstrate that it improves generalization on the sequence-clustered SKEMPI2 and enables models not originally designed for this task, such as ProSST Li et al. (2024b) and the ESM family Hayes et al. (2024), to reach or surpass state-of-the-art specialized PPI predictors.

Limitations of the SKEMPI2 Dataset

Latent Sequential and Structural Data Leakage

The main issue with the SKEMPI2 dataset is systematic data leakage. Prior studies split training and test sets by complex PDB ID Luo et al. (2023); Yu et al. (2024); Mo et al. (2024); Wee and Wei (2024), but many differently

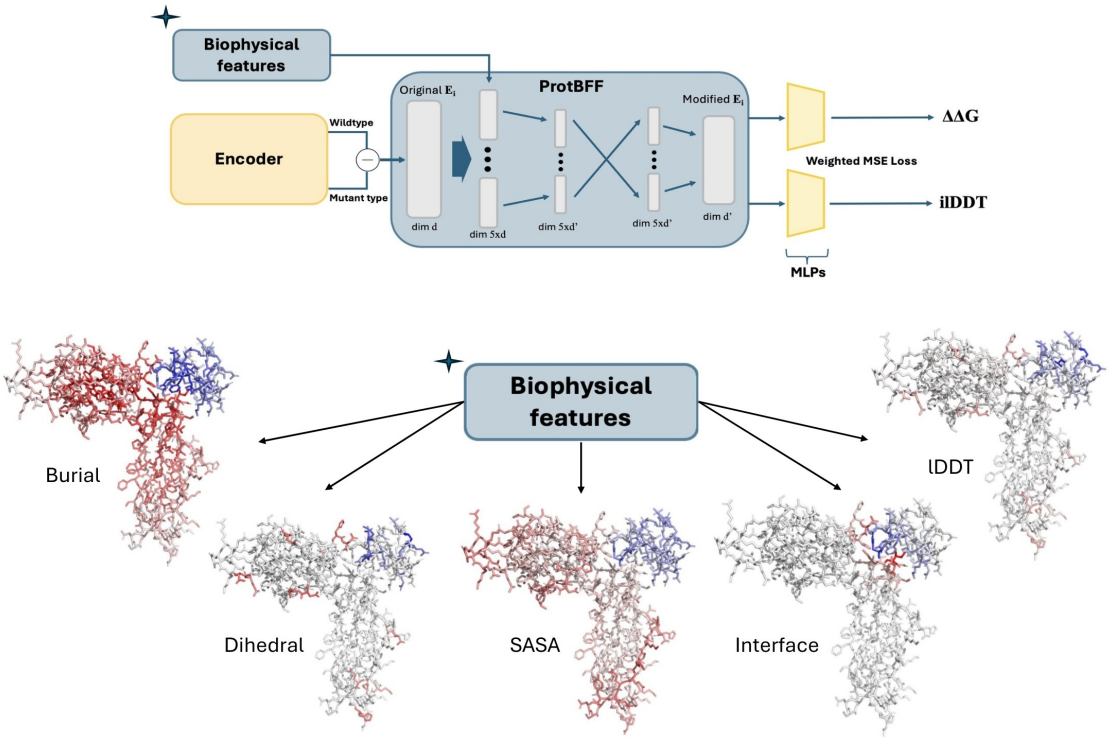

Figure 1. The upper part of the figure shows a simplified schematic of the ProtBFF framework. The pipeline begins with embedding extraction, followed by embedding subtraction and scaling using biophysical features. These processed embeddings are then passed through a pooling layer, a cross-embedding attention module, and finally two MLP heads to generate predictions of $\Delta\Delta G$ and ilDDT, optimized via a weighted loss function. The lower part of the figure illustrates the five biophysical features by coloring residues according to their scores, with chain A shown in red and chain B in blue. Residues with higher biophysical contributions are colored more intensely, while those with weaker contributions appear closer to gray.

named complexes are highly homologous in sequence and structure Bushuiev et al. (2024); Jankauskaitė et al. (2018); Yu et al. (2025) (Figure 2 B). As a result, homologous proteins often appear in both splits, inflating performance.

To quantify this effect, we adopted a stricter dataset splitting strategy based on sequence similarity. Specifically, we used the CD-HIT clustering algorithm Li and Godzik (2006) to group protein complexes by sequence identity, applying thresholds from 60% to 100%. Two complexes are clustered together if the chain sequences exceed the specified similarity threshold. At 100% identity, this procedure is equivalent to no clustering, producing one cluster per complex. As the identity threshold is lowered, the number of clusters decreases rapidly: As shown in Table 1 at 60% identity, the dataset is reduced to 136 clusters, underscoring the strong homology within SKEMPI2. Further reductions below 60% lead to a rapid consolidation into only one primary cluster. Notably, reducing the identity threshold from 100% to 99% decreases the number of clusters from 335 to 253, indicating substantial redundancy in SKEMPI2 and showing that clustering by complex name alone is insufficient to prevent overlap. Table 1 summarizes the number of clusters across thresholds.

Beyond sequence similarity, SKEMPI2 aggregates

Threshold	60%	80%	95%	99%	100%
Clusters	136	179	219	253	335

Table 1. Number of protein complex clusters at different sequence identity thresholds. Lower thresholds merge homologous complexes, highlighting the redundancy in SKEMPI2.

$\Delta\Delta G$ measurements obtained using diverse experimental techniques with varying levels of reliability Jankauskaitė et al. (2018), which introduces additional noise into the dataset. As shown in Figure 2 C, model performance depends strongly on the experimental method used for measurements. In particular, indirect or surface-based assays, such as Enzyme-Linked Immunosorbent Assay (ELISA), are less reliable for quantitative binding-affinity estimation and can introduce substantial measurement noise. This increased experimental variance degrades predictive accuracy and makes the corresponding data more difficult for models to learn from.

Impact of Data Leakage on Models' Performance

To accurately assess the impact of the aforementioned data leakage on the performance of models predicting $\Delta\Delta G$, we benchmarked four widely used deep learning models: DDAffinity Yu et al. (2024), ProMIM Mo et al. (2024), RDE-Network Luo et al. (2023), and

ProSST Li et al. (2024b), using 10-fold cross-validation on the aforementioned sequence-similarity–clustered SKEMPI2 dataset. As the sequence-identity threshold is lowered, the resulting train–test splits contain progressively fewer homologous proteins; consequently, model performance is expected to decrease if previously reported results were inflated by redundancy or information leakage. This evaluation therefore provides a systematic test of model robustness to homology-aware data splitting.

As shown in Figures 2 A, while models achieve high Pearson correlations of above 0.6 and Spearman correlations of above 0.5 when no clustering (the 100% threshold) is applied, performance declines consistently and substantially as the threshold is reduced. This trend highlights that previously reported results on SKEMPI2 were strongly influenced by redundancy between training and test sets, ultimately leading to overfitting.

Results

Prediction Improvements with ProtBFF

To assess the improvements provided by ProtBFF as a plug-in module, we integrated it into two types of models that follow the latent representation paradigm and benchmarked them against several established baselines: ProSST, originally developed for single-protein stability prediction Li et al. (2024b), and the ESM2 Lin et al. (2023) and ESM3 models Hayes et al. (2024), two general-purpose protein language models (the first one sequence-only, and the other integrating sequences and structures). All models were retrained using a 10-fold cross-validation protocol on the SKEMPI2 dataset clustered at 60% sequence identity, as detailed in the Methods section

ProSST Li et al. (2024b), though not originally designed for protein complex $\Delta\Delta G$ prediction, improves substantially with ProtBFF, with the Pearson increasing from 0.428 to 0.514 and the Spearman increasing from 0.354 to 0.477, surpassing specialized models such as ProMIM and DDAffinity. ESM models also benefit substantially, with Pearson and Spearman correlations increasing to levels comparable to most state-of-the-art models.

These results illustrate that models pretrained on large protein datasets, such as the CATH database Knudsen and Wiuf (2010) for ProSST and the extensive sequence and structural corpora underlying ESM2 and ESM3, learn rich general-purpose representations Hayes et al. (2024); Li et al. (2024b). By guiding these embeddings with ProtBFF, which selectively amplifies structurally relevant residues through biophysical features, we are able to unlock performance that rivals or even surpasses methods specifically designed for protein complex $\Delta\Delta G$ prediction. Further information about the benchmarking and integration process can be found in the Methods section.

We also evaluated the effect of applying ProtBFF across the ESM2 model family, ranging from 150 million to 15 billion parameters. With ProtBFF, the 150-million-parameter model notably outperforms even much larger counterparts, as well as all other standard ESM2 variants. ProtBFF also provides consistent performance gains across all model sizes (Figure 3). Interestingly, the standard the ESM2 650M model outperforms the 3B and 15B variants on SKEMPI2, consistent with observations that smaller ESM models sometimes match or exceed larger ones on specific predictive tasks Vieira et al. (2025).

Ablation Study Results. To assess the contribution of the five biophysical scores and the auxiliary iIDDT loss to ProtBFF's performance, we conducted an ablation study. Each component was systematically removed in turn, and we observed that all scores as well as the iIDDT loss provided measurable improvements in both Pearson and Spearman correlations. Among them, the interface and burial features contributed the largest gains. These findings demonstrate that ProtBFF's accuracy arises from the integration of multiple complementary biophysical signals rather than reliance on any single feature.

Table 2 presents the effects of systematically removing each biophysical feature from ProtBFF. All features contribute positively to predictive performance, with the interface and residue scores showing the largest impact; removing these two features leads to substantial drops in both Pearson and Spearman correlations, confirming their central role in capturing mutational effects on binding affinity. The dihedral, SASA, and lDDT scores produce smaller, yet still meaningful, reductions in performance, indicating that they provide complementary structural information that refines the model's predictions. Additionally, removing the iIDDT-based loss function results in a measurable decline in both correlation metrics, demonstrating that incorporating structural fidelity into the training objective improves model robustness. Finally, a baseline variant without any biophysical scores performs significantly worse, highlighting that the multi-feature integration strategy is crucial for ProtBFF's predictive power.

SARS-CoV-2 Dataset Results. We next tested how well ProtBFF-enhanced models generalize to Out-of-Distribution (OOD) tasks, focusing on $\Delta\Delta G$ prediction for virus-receptor and antibody–antigen binding, which are key to antibody design and pandemic preparedness yet also associated with potential biosecurity risks Feldman and Feldman (2025); Wang et al.; Yin and Pierce (2024). Performance was evaluated on a DMS dataset measuring SARS-CoV-2 RBD binding to ACE2, LY-CoV555, and REGN10987 Starr et al. (2020, 2021). Pretrained ProSST- and ESM2-based models, with and without ProtBFF, were assessed under zero-shot and few-shot learning across varying training set sizes, as shown in Figures 4 A–C.

This expanded benchmarking yields two key insights.

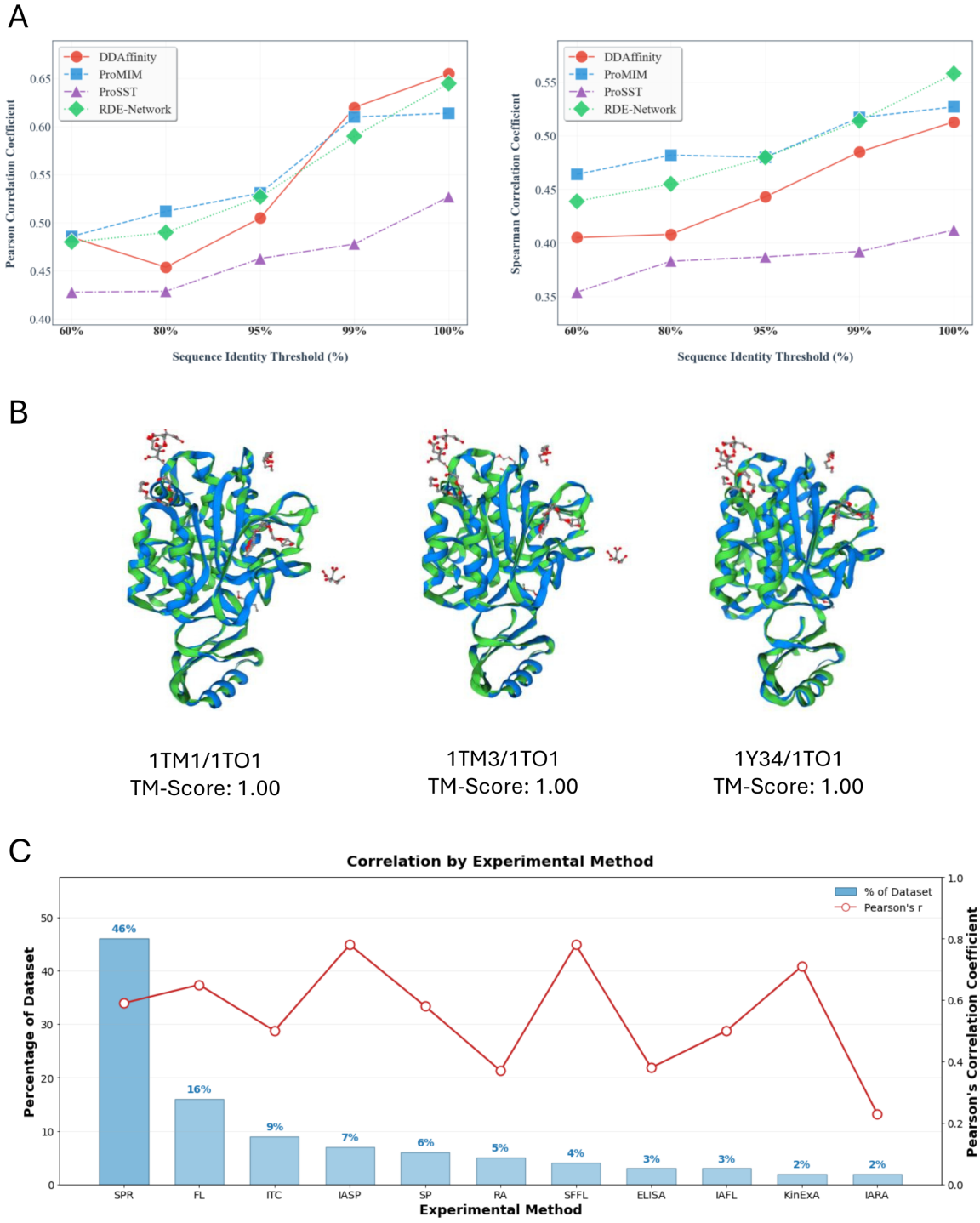

Figure 2. A Performance of deep learning $\Delta\Delta G$ predictors on SKEMPI2 under progressively stricter sequence-identity clustering. Lines report correlation for each model, with the left subfigure being the Pearson correlation and the right being the Spearman correlation. Model performance declines sharply as homology constraints are enforced. **B** Structural superposition of three representative SKEMPI2 protein complexes (1TM1, 1TM3, and 1Y34) onto a fourth complex using US-Align Zhang et al. (2022). Although treated as distinct dataset entries, these complexes share over 99% sequence identity and exhibit near-identical structures, illustrating the extent of hidden redundancy in SKEMPI2. **C** Leave-one–experimental-method-out evaluation of the $\Delta\Delta G$ predictor DDAffinity on SKEMPI2. For each experimental method, the model was trained on all datapoints except those measured using the held-out method and tested exclusively on that method. Only experimental methods comprising at least 1% of the dataset are shown. To isolate effects attributable to experimental methodology, homology-based clustering was not applied in this analysis. Pearson's correlation coefficient is shown on the y-axis, with experimental methods on the x-axis.

First, ProtBFF consistently improves both ProSST and ESM2 representations when paired with a simple classifier. Across datasets, ProSST with ProtBFF often outperforms its base version by over 0.2 in Pearson and Spearman correlations, while ESM2 gains even more, approaching the performance of ProSST with ProtBFF.

Second, although all models perform poorly in the zero-shot setting because SKEMPI2 provides limited cov-

Benchmarking and Ablation Results	Baseline		With ProtBFF	
Method / Variant	Pearson (ρ)	Spearman (r)	Pearson (ρ)	Spearman (r)
Model Benchmarking				
ProSST Li et al. (2024b)	0.428	0.354	**0.514**	**0.477**
ESM2 Lin et al. (2023)	0.194	0.204	0.451	0.410
ESM3 Hayes et al. (2024)	0.159	0.099	0.362	0.347
ProMIM Mo et al. (2024)	0.486	0.464	*	*
RDE-Network Luo et al. (2023)	0.480	0.439	*	*
DDAffinity Yu et al. (2024)	0.485	0.405	*	*
RDE-Linear Luo et al. (2023)	0.369	0.360	*	*
FoldX Schymkowitz et al. (2005)	0.320	0.294	*	*
Ablation Study on ProSST				
None (Full ProSST+ProtBFF)	–	–	**0.514**	**0.477**
Interface Score removed	–	–	0.462	0.418
Burial Score removed	–	–	0.471	0.426
Dihedral Score removed	–	–	0.498	0.458
SASA Score removed	–	–	0.503	0.464
lDDT Score removed	–	–	0.506	0.469
ilDDT Loss Function removed	–	–	0.505	0.459
All Scores removed	–	–	0.436	0.385

Table 2. Benchmarking and ablation results. Top: Performance of baseline encoders with and without ProtBFF.
Bottom: Ablation study of ProtBFF applied to ProSST, where each row removes one feature or the auxiliary *ilDDT* loss. Values reported are Pearson (ρ) and Spearman (r) correlations. Bold indicates best performance. Asterisks (*) mark methods where ProtBFF is not applicable.

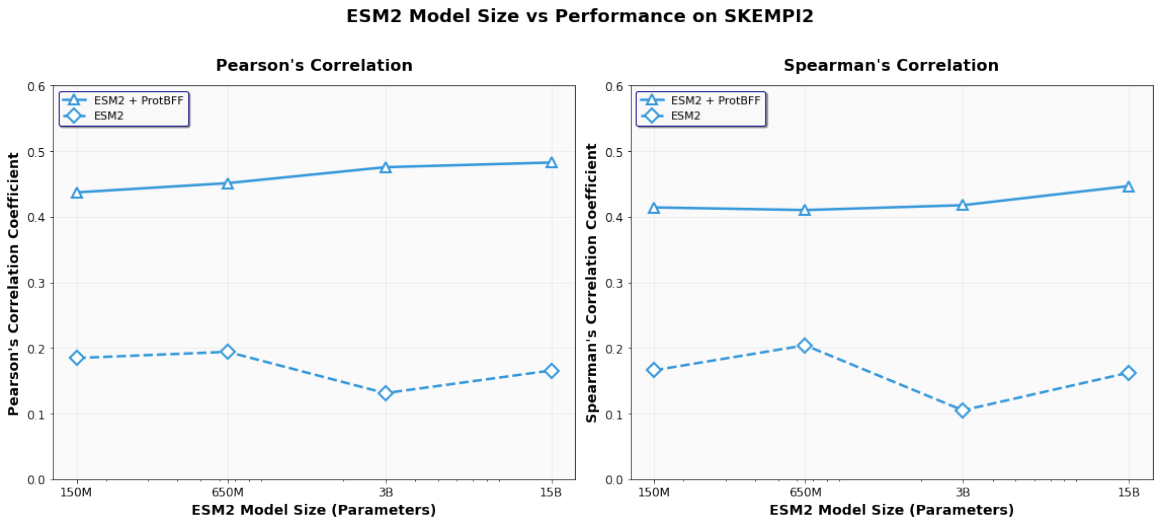

Figure 3. Performance of ESM2 on the SKEMPI2 dataset measured by Pearson's and Spearman's correlation coefficients across model sizes, defined by the number of parameters.

erage of antibody–antigen and virus–receptor interactions, few-shot learning markedly improves accuracy. With only 10% of training data, ProtBFF-enhanced models achieve high predictive performance, suggesting practical utility in data-limited contexts such as active learning.

Notably, the strong fine-tuned performance reflects the dataset's structure. Evaluations are on DMS data from a single protein complex, making prediction easier than cross-complex generalization. Repeated exposure to mutations on the same backbone allows the models to leverage ProtBFF's biophysical features more effectively, enhancing embeddings and boosting performance metrics.

Discussion

Predicting mutational effects on protein–protein binding highlights a central challenge in machine learning for biology. Models must generalize reliably from datasets that are small, biased, and unevenly distributed. Although self-supervised pretraining on massive protein sequence databases has transformed structure prediction and representation learning Abramson et al. (2024); Hayes et al. (2024); Watson et al. (2023), supervised $\Delta\Delta G$ prediction remains fundamentally data-limited. Benchmarks such as SKEMPI2 contain only 7,085 measurements across 345 complexes Pucci et al. (2022); Bushuiev et al. (2024); Jankauskaitė et al. (2018); Deng et al. (2025), restricting the ability of models to learn transferable biophysical principles.
Our analysis reveals that these constraints are exacer-

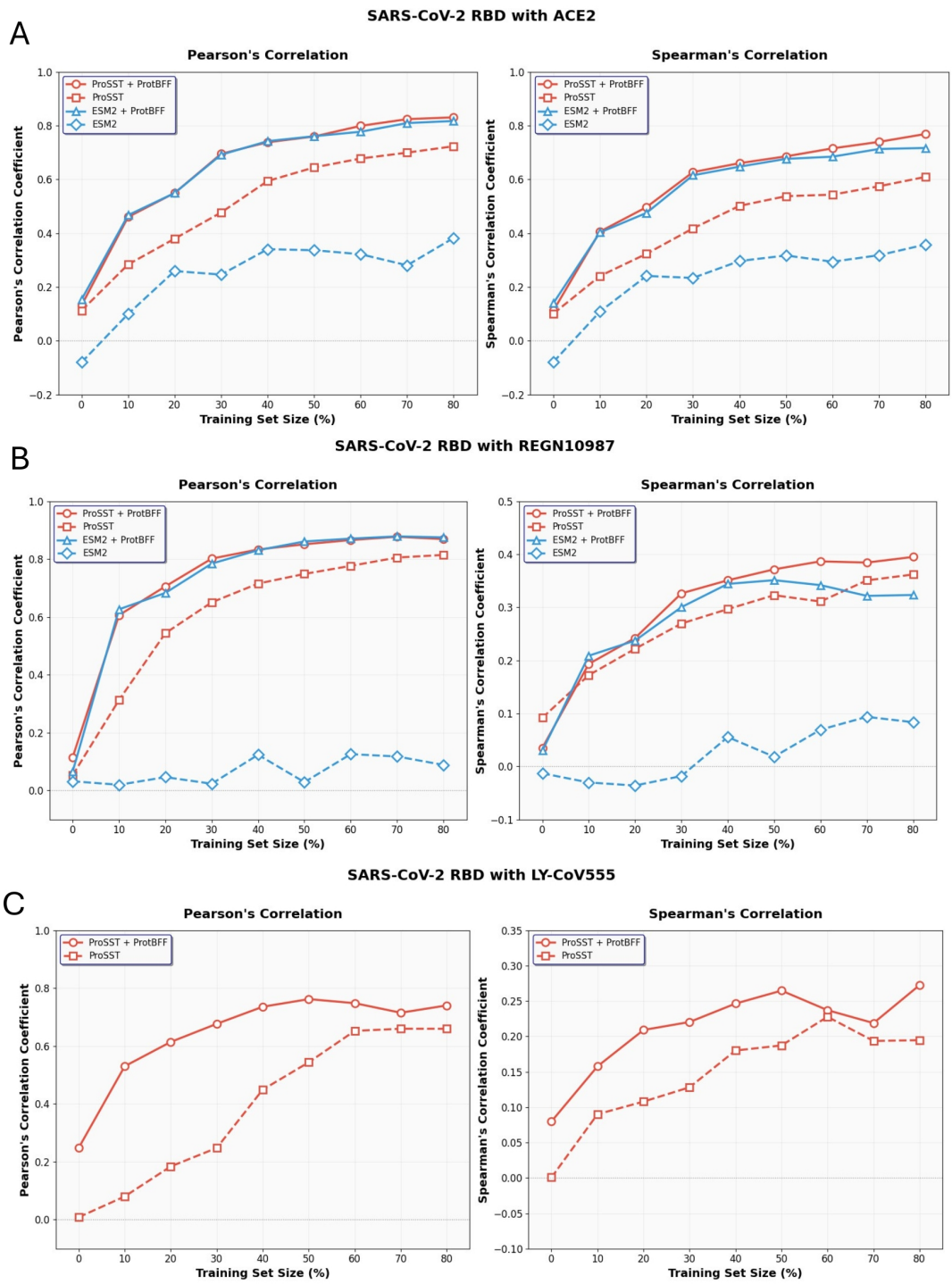

Figure 4. A–C Performance of ProSST and ESM2, with and without ProtBFF, on deep mutational scanning (DMS) datasets of SARS-CoV-2 RBD mutations when bound to (A) ACE2, (B) REGN10987, and (C) LY-CoV555. The left column reports Pearson's correlation coefficient, while the right column reports Spearman's correlation coefficient. *Note:* Due to the large size of the RBD–LY-CoV555 complex template structure (PDB 7KMG), it exceeds the context length limit of ESM2 and therefore could not be processed. As a result, ESM2-based results are omitted from subfigure C.

bated by widespread data leakage resulting from sequence and structural redundancy. Train–test splits based solely on complex identifiers fail to control for homology, leading to inflated performance estimates and overstated generalization. When SKEMPI2 is partitioned using sequence-identity–based clustering, predictive accuracy drops sharply across methods, indicating that many models rely in part on memorizing dataset-specific patterns rather than learning biophysical mechanisms. Similar issues have been observed

in other machine learning domains, where overlap between training and evaluation data leads to misleading assessments of model capability Roelofs et al. (2019); Jiralerspong et al. (2023).

These results underscore the importance of benchmarks and modeling strategies that explicitly value mechanistic understanding. Pretrained embeddings are powerful, but when used in isolation, they are insufficient in low-data regimes. Instead, they should be guided by physically interpretable, task-relevant priors that constrain learning toward biophysical regularities.

ProtBFF consistently improves performance across encoder architectures by integrating biophysical information directly into the learning process. Ablation experiments show that predictive gains arise from combining multiple sources of physical signal rather than from any single feature. Interface- and residue-level descriptors contribute the most, reflecting their direct relevance to binding affinity, while global structural features such as solvent accessibility, dihedral geometry, and lDDT provide smaller but consistent improvements by adding complementary context.

At the same time, ProtBFF should be viewed as an initial step rather than a complete solution. The five biophysical features used here represent a coarse approximation of the physical processes governing binding. In particular, FoldX models mutations through side-chain substitutions on a fixed backbone, which prevents the representation of backbone relaxation, conformational coupling, and dynamic effects Schymkowitz et al. (2005); Alford et al. (2017). Important aspects of protein energetics, therefore, remain unmodeled, and incorporating richer physical information will be essential for further progress.

ProtBFF also relies on encoder models that produce residue-level embeddings, enabling alignment between learned representations and biophysical features. More complex architectures that operate on higher-order geometric or relational representations may require specialized mechanisms for integrating physical priors directly into the model design Yu et al. (2024); Marquet et al. (2022). This highlights the broader challenge of co-designing architectures and inductive biases rather than treating physical information as a modular add-on.

Beyond model design, our findings underscore the importance of rigorous dataset construction and evaluation. Without careful control for redundancy, increasingly sophisticated models risk becoming effective memorization systems with limited practical relevance Graber et al. (2025). Developing larger, more diverse, and experimentally robust datasets is therefore critical for translating methodological advances into reliable predictive tools.

The implications extend beyond benchmark performance. Accurate and generalizable $\Delta\Delta G$ predictors are essential for identifying beneficial mutations in therapeutic antibodies, guiding protein engineering, and anticipating the evolution of viral variants Yin and Pierce (2024); Thadani et al. (2023); Marquet et al. (2022). In these settings, models must reflect biophysical reality rather than exploit dataset artifacts. Taken together, our results suggest that progress in $\Delta\Delta G$ prediction will depend on tightly integrating machine learning with biophysical reasoning. ProtBFF demonstrates that even relatively simple physical priors can meaningfully improve generalization, while also highlighting the need for deeper and more comprehensive representations of protein physics.

Materials and Methods

Dataset Collection

SKEMPI Dataset Preprocessing. In this study, we used the SKEMPI2 dataset for both benchmarking and training. SKEMPI2 contains 345 protein complexes and 7,085 combinatorial point mutations with experimentally determined $\Delta\Delta G$ values, with structures obtained from the Protein Data Bank (PDB) Jankauskaitė et al. (2018). Due to input size limits of the smallest benchmarked model, ProSST (which was not trained to encode protein complexes), the 10 largest complexes were excluded Li et al. (2024b). This yielded a final working set of 335 complexes and 6,631 mutational variants.

To ensure fair comparisons, this reduced SKEMPI2 dataset was used consistently for all training and benchmarking. We applied ten-fold cross-validation using identical splits across all experiments. Before generating the splits, we clustered sequences with CD-HIT to ensure that complexes in different folds shared less than a specified sequence homology threshold Li and Godzik (2006). This step mitigated the data leakage issue in SKEMPI, since high sequence similarity often implies structural similarityKrissinel (2007); Koehl and Levitt (2002). Given that SKEMPI2 contains many complexes with nearly identical composition, this clustering also reduced structural homology overlap across folds Jankauskaitė et al. (2018).

FoldX Utilization. FoldX was used to generate mutant protein structures for all complexes in the SKEMPI2 dataset. After applying RepairPDB twice to the wildtype structures, FoldX relaxes the mutant side chains using the BuildModel function Schymkowitz et al. (2005). Importantly, this procedure does not modify the protein backbone. Although not explored here, alternative methods that incorporate backbone relaxation, such as Rosetta Relax, may produce more accurate mutant structures (albeit at higher computational cost) which could in turn yield improved biophysical features and enhance $\Delta\Delta G$ prediction accuracy Alford et al. (2017).

ProtBFF architecture

A central paradigm in molecular biology is to learn expressive latent representations of molecules and their interactions, while using relatively simple models for downstream prediction tasks such as $\Delta\Delta G$ Marquet

et al. (2022); Li et al. (2024a); Dahal et al. (2025). Concretely, an encoder generates per-residue embeddings for both wildtype and mutant proteins, integrating sequential and or structural information Mo et al. (2024); Luo et al. (2023); Yu et al. (2024). These encoders are typically pretrained on large-scale datasets with self-supervised objectives (e.g., masked residue prediction) to build informative internal representations Marquet et al. (2022); Li et al. (2024b). In this study, we propose a general biophysical framework that augments such models by integrating explicit biophysical features into per-residue embeddings, which can be seen in Figure 1.

Our framework builds on the observation that embedding-based models produce a latent representation for every residue, yet not all residues contribute equally to changes in binding free energy. Prior studies have shown that simple per-residue biophysical scores can strongly correlate with mutational effects on binding affinity Yu et al. (2024); Dauparas et al. (2022). Motivated by this, we enrich residue embeddings by scaling them with biophysical metrics: embeddings of residues with high structural or interfacial relevance are amplified, while those less likely to contribute are down-weighted. Incorporating multiple complementary scores enables the model to integrate diverse structural cues, while cross-embedding attention allows these signals to interact and emphasize the most informative patterns Girdhar and Ramanan (2017); Tang et al. (2023). Together, this yields richer and more interpretable residue-level representations.

To implement this idea, we generate five scaled copies of each residue embedding, each weighted by a distinct biophysical metric: interface propensity, residue burial, dihedral deviation, solvent-accessible surface area (SASA), and local distance difference test (lDDT), which are visualized in Figure 1. These metrics are computed from the wildtype structures and FoldX-generated mutant structures Schymkowitz et al. (2005). Formally, for residue i with embedding Given $\mathbf{E}_i \in \mathbb{R}^M$, we construct

$$\mathbf{E}_i^{(k)} = s_i^{(k)} \cdot \mathbf{E}_i, \quad k \in \{\text{interface}, \text{burial}, \text{dihedral}, \text{SASA}, \text{lDDT}\}.$$

where $s_i^{(k)} \in [0,1]$ is the biophysical score for metric k. The set $\{\mathbf{E}^{(k)}\}_{k=1}^5$ constitutes the enriched embedding space used in the downstream tasksDing et al. (2021). As illustrated in Figure 1, each scaled embedding is projected into a lower-dimensional space, regularized through dropout and non-linear activation. A cross-embedding multihead attention mechanism then integrates the five streams, allowing the model to reweight and combine information across biophysical perspectives.Ding et al. (2021); Xu et al. (2019) Finally, an attention pooling layer aggregates these signals into a compact representation. This design introduces minimal parameters, making it well-suited for limited datasets such as SKEMPI2. Please refer to Methods for more details.

To further guide learning, we impose a multi-task weighted loss that conditions the network to recover not only experimental $\Delta\Delta G$ values but also the interface structural consistency metric iIDDT Mariani et al. (2013); Thomas et al. (2025). This auxiliary prediction encourages the extraction of structurally meaningful features and improves generalization.

Overall, ProtBFF can be applied as a drop-in replacement for the feed-forward modules placed after protein embeddings in $\Delta\Delta G$ prediction pipelines. Because it requires only minimal adaptation to embedding dimensionality, it can readily be integrated with a wide range of pretrained encoders and extended to other downstream structural prediction tasks, offering a general and practical means of injecting biophysical inductive bias into deep learning models.

Embedding-aware attention Network Architecture. The ProtBFF framework operates on embeddings generated by a pre-trained encoder. For each protein complex, 5 per-residue biophysical scores are computed, each associated with their own scaled embedding. The k-th of the five embeddings is denoted by $\mathbf{E}^{(k)} \in \mathbb{R}^{L\times M}$, where L is the number of residues in the protein complex and M is the dimensionality of the embeddings. L corresponds to the full sequence length of the complex.

Although the specific embedding dimensionality M may differ across different encoders, the embedding-aware attention network is agnostic to this and can be adapted by adjusting its input size hyperparameter. All embeddings are of $\mathbb{R}^{L\times M}$, thereby allowing for per-residue scaling and subsequent pooling operations.

Once the per-score embeddings are obtained, they are max-pooled along the residue dimension, producing five fixed-length vectors in $\mathbb{R}^M$. These are concatenated to form a single representation $\mathbf{X} \in \mathbb{R}^{5M}$, which is passed into the embedding-aware attention network. The network first reshapes $\mathbf{X}$ back into its 5 constituent $\mathbb{R}^M$ embeddings, which then each undergo an attention pooling operation. The resulting 5 pooled vectors are processed via a cross-embedding attention mechanism, which attends over all embeddings simultaneously, and normalized to produce a single $\mathbb{R}^{512}$ representation. This final representation is passed through a three-layer multilayer perceptron (MLP) to produce $\Delta\Delta G$ predictions. The reported value is the average of the predictions for the forward and reverse mutations to preserve anti-symmetry:

$$\widehat{\Delta\Delta G} = \frac{f(\mathbf{X}_{\text{forward}}) - f(\mathbf{X}_{\text{reverse}})}{2}.$$

For a full schematic of the ProtBFF framework, please refer to Figure 1

It is possible to fine-tune the model architecture based on the amount of data or the specific task, while maintaining the general methodology—namely, the integration of biophysical features followed by attention-pooling.

Definition of Biophysical Scores

We define five novel biophysical scores, which are generated from the structural conformation information, as follows.

The ***interface score*** is a normalized metric that quantifies how close a given residue is to a protein-protein interface within a protein complex. A higher interface score indicates that the residue is closer to the protein-protein interface and plays a more significant role in inter-chain interactions, and, therefore, a mutation at that residue would be more influential on structural stability Zhao et al. (2025); Mirabello and Wallner (2024).

The interface score for residue i is

$$\tilde{I}_i = \frac{\sum_{j \in N(i,9)} e^{-d_{ij}^2/(2\sigma_{\text{interface}}^2)}\, C_{\text{inter}}(j)}{\max_m I_m},$$

where $N(i,9)$ are the 9 nearest neighbors (including i), d_{ij} is the residue–residue distance, and $\sigma_{\text{interface}}$ controls the Gaussian width. The inter-chain contact count is

$$C_{\text{inter}}(j) = \sum_{\ell \neq j} 1\big(d_{ij} < 10\,\text{Å} \wedge \text{chain}(\ell) \neq \text{chain}(i)\big)\,,$$

and $\max_m I_m$ is the maximum raw score (the numerator) across all residues. This normalization yields values between 0 and 1 for direct comparison across residues.

The ***burial score*** quantifies how deeply buried a given residue is within a protein, informed by the intuition that mutations to more buried residues generally cause more significant conformational changes Yu et al. (2024). A higher burial score indicates a residue is more embedded in the protein's interior.
The burial score for residue i is

$$c_i = \frac{\sum_{j \in N(i,9)} c_j}{\max_{t \in L} \sum_{j \in N(t,9)} c_j},$$

where $N(i,9)$ are its 9 nearest neighbors and c_j is the count of residues within 10 Å of residue j. Normalizing by the maximum neighbor-sum over all residues L yields values between 0 and 1.

The ***dihedral score*** quantifies structural changes in side-chain dihedral (chi) angles after mutation. Mutations causing significant chi angle alterations can lead to substantial conformational rearrangements, impacting protein stability and function. A higher dihedral score indicates a more pronounced structural shift.
The dihedral score for residue i is computed by comparing side-chain dihedral angles between the wild-type and optimized structures. Specifically, for each residue $j \neq i$ and each of its side-chain dihedral angles χ_k, we compute the angular deviation as

$$\Delta\chi_{j,k} = \min\left(\left|\chi_{j,k}^{\text{opt}} - \chi_{j,k}^{\text{wt}}\right|, 360^\circ - \left|\chi_{j,k}^{\text{opt}} - \chi_{j,k}^{\text{wt}}\right|\right).$$

where $\chi_{j,k}^{\text{wt}}$ and $\chi_{j,k}^{\text{opt}}$ denote the k-th side-chain dihedral angle (χ) of residue j in the wild-type and optimized structures, respectively. This definition selects the minimum angular distance while accounting for angular periodicity. Symmetric 180° flips in aromatic residues (PHE, TYR, HIS) are treated as equivalent conformations and are therefore ignored.
The total angular change associated with residue i is obtained by summing over all other residues:

$$\Delta_{\text{total}}(i) = \sum_{j \neq i} \sum_{k=1}^{n_j} \Delta\chi_{j,k},$$

where n_j is the number of side-chain dihedral angles for residue j. The dihedral score is then normalized by the maximum total angular change across all residues:

$$\text{Dihedral}_{\text{score}}(i) = \frac{\Delta_{\text{total}}(i)}{\max_j \Delta_{\text{total}}(j)}.$$

The ***SASA score*** quantifies the degree to which a residue is exposed to solvent in a protein structure, with higher values indicating greater solvent exposure Gao et al. (2021). This score is calculated using the *freesasa* package, which determines the Solvent Accessible Surface Area of each residue based on atomic coordinates in the wildtype structure Mitternacht (2016). The SASA score incorporates both the solvent exposure of the residue of interest and the exposure of its nearby residues, weighted by their sequential proximity.
The SASA score for residue i is the Gaussian-weighted sum of solvent-accessible surface areas for its 10 nearest sequential neighbors, normalized to $[0,1]$:

$$\text{SASA}_i = \frac{\sum_{j \in N(i,10)} \text{SASA}_j\, w_{ij}}{\max_{t \in \{L\}} \sum_{j \in N(t,10)} \text{SASA}_j\, w_{tj}}$$

Here, SASA_j is the solvent-accessible surface area (from *freesasa*) of neighbor j, $w_{ij} = e^{-d_{ij}^2/(2\sigma_{\text{SASA}}^2)}$ is a Gaussian weight favoring closer neighbors, $N(i,10)$ is the 10 nearest sequential neighbors to the left and right of i, and $\{L\}$ is the set of all residues.

The ***IDDT score*** measures per-residue atomic conformational changes between the wildtype (template) and FoldX-generated mutant (predicted) structures Mariani et al. (2013). Larger changes indicate residues more likely to affect overall structure and function. The score is normalized to $[0,1]$, and we scale embeddings by $1 - \text{IDDT}$ so that residues with greater changes receive higher emphasis during max pooling.

Attention-model Training and Evaluation. The embedding-aware attention network is trained separately, and, for the SKEMPI2 dataset, it is trained on

embeddings generated from the wildtype and mutant sequences and structures present in the dataset. The network contains two multilayer perceptrons (MLPs): one for predicting $\Delta\Delta G$ and one for predicting interfacial lDDT (ilDDT). The ilDDT metric is equivalent to lDDT but restricted to interchain contacts, producing a single scalar value for the entire protein complex rather than a per-residue score Mariani et al. (2013); Biasini et al. (2013). The ilDDT-predicting MLP is included during training as a regularization mechanism, encouraging the network to learn features relevant to structural accuracy in addition to energetic changes.
Because the embeddings are highly feature-rich, conditioning the model to recover ilDDT promotes the extraction of meaningful structural information. This joint training is implemented using a modified mean squared error (MSE) loss:

$$\mathcal{L} = 1.0 \cdot \mathrm{MSE}(\widehat{\Delta\Delta G}, \Delta\Delta G) + 0.2 \cdot \mathrm{MSE}(\widehat{\mathrm{ilDDT}}, \mathrm{ilDDT}),$$

where $\widehat{\Delta\Delta G}$ and $\widehat{\mathrm{ilDDT}}$ are the network predictions, and the ground-truth ilDDT is calculated by comparing the FoldX-mutated structure to the wildtype PDB template. The use of two output heads ensures that conditioning occurs early in the network, at the cross-embedding attention stage, after which the $\Delta\Delta G$ head can specialize in energetic predictions Girdhar and Ramanan (2017); Lee et al. (2019). During inference, the ilDDT output is ignored, as it serves only as a conditioning signal during training.

$\Delta\Delta G$ Predictor Benchmarking

We benchmarked eight baseline models, several of which are regarded as state-of-the-art, and evaluated all models using ten-fold cross-validation. For ESM3, both sequence and structures were embedded Hayes et al. (2024), while for ESM2, only sequences were embedded Lin et al. (2023). Although ProMIM and DDAffinity incorporate features and code from RDE-Network, their architectures differ substantially and have been reported to outperform it, making the comparison particularly relevant Mo et al. (2024); Luo et al. (2023); Yu et al. (2024).

Model Training and Testing. Benchmarking for consistent comparison across all models was performed on the 60% sequence similarity split of the SKEMPI2 dataset. The models were benchmarked on the entire dataset, including both single and multiple point mutations.
Models not requiring retraining on SKEMPI2 (namely ESM2, ESM3, and ProSST) were used out-of-the-box, while those requiring retraining (ProMIM, RDE-Network, RDE-Linear and DDAffinity) were retrained across all ten folds according to their respective protocols. When ProSST, ESM2, and ESM3 were adapted to the ProtBFF framework, the encoders were not retrained; instead, their embeddings were used to train the attention network for $\Delta\Delta G$ prediction.

For the general benchmarking, the ESM2 model used was esm2_t33_650M_UR50D. For the specific benchmarking across ESM2 model sizes, the models evaluated were esm2_t30_150M_UR50D, esm2_t33_650M_UR50D, esm2_t36_3B_UR50D, and esm2_t48_15B_UR50D.

Ablation Study on Biophysical Feature Contributions. To evaluate the relative contributions of the five biophysical features used in ProtBFF, we performed a systematic ablation study. Each variant of the model was trained with one or more features removed at a time, while keeping the remaining components intact. This allowed for the quantification of their individual and collective importance in driving $\Delta\Delta G$ prediction performance.
We used the ProSST backbone as the base model for this analysis and trained each ablated ProtBFF variant on the SKEMPI2 dataset clustered at 60% sequence identity, following the same 10-fold cross-validation protocol described in section* .

SARS-CoV-2 Benchmarking. For further benchmarking, we used the DMS dataset from Bloom paper, which quantifies changes in binding affinity between the SARS-CoV-2 RBD, the mutational target, and three binding partners: the human ACE2 receptor, and the neutralizing antibodies LY-CoV555 and REGN10987 Starr et al. (2021). The corresponding template structures used for these complexes were PDB IDs 7W9I (RBD–ACE2), 7KMG (RBD–LY-CoV555), and 9LYP (RBD–REGN10987). Although additional antibodies are present in the dataset, they were excluded because their complex sizes exceed the context window limits of ESM and ProSST, or because their corresponding template structures are sparsely represented in the PDB, preventing reliable generation of structural embeddings. After removing entries with missing measurements and filtering out mutations located in structurally unresolved regions, the final datasets comprised 3,669 datapoints for the RBD–ACE2 interaction, 3,156 datapoints for RBD–REGN10987, and 3,137 datapoints for RBD–LY-CoV555. The aforementioned PDB structures were used as wild-type templates, and all mutations were introduced using FoldX following the same protocol applied to SKEMPI2.
To benchmark performance on this dataset, we loaded the models trained during the SKEMPI2 benchmarking. Specifically, models from the first fold of the ten-fold cross-validation were selected for consistency. These pretrained models were then progressively fine-tuned on the Bloom DMS datasets using increasing training set sizes, following the protocol described earlier and illustrated in Figure 4.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A. I., Cowie, A., Figurnov, M., Fuchs, F. B., Gladman, H., Jain, R., Khan, Y. A., Low, C. M. R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E. D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D., and Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500. doi: 10.1038/s41586-024-07487-w.

Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L. J., Das, R., Baker, D., Kuhlman, B., Kortemme, T., and Gray, J. J. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation*, 13(6):3031–3048. doi: 10.1021/acs.jctc.7b00125. Publisher: American Chemical Society.

Biasini, M., Schmidt, T., Bienert, S., Mariani, V., Studer, G., Haas, J., Johner, N., Schenk, A. D., Philippsen, A., and Schwede, T. (2013). OpenStructure: an integrated software framework for computational structural biology. *Acta Crystallographica Section D: Biological Crystallography*, 69(5):701–709. doi: 10.1107/S0907444913007051. Publisher: International Union of Crystallography.

Bushuiev, A., Bushuiev, R., Kouba, P., Filkin, A., Gabrielova, M., Gabriel, M., Sedlar, J., Pluskal, T., Damborsky, J., Mazurenko, S., and Sivic, J. Learning to design protein-protein interactions with enhanced generalization, (2024). URL https://arxiv.org/abs/2310.18515.

Dahal, A., Murad, S. A., and Rahimi, N. Embedding Shift Dissection on CLIP: Effects of Augmentations on VLM's Representation Learning, (2025). URL http://arxiv.org/abs/2503.23495. arXiv:2503.23495 [cs].

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., de Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., King, N. P., and Baker, D. (2022). Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56. doi: 10.1126/science.add2187.

Deng, A., Householder, K., Wu, F., Thrun, S., Garcia, K. C., and Trippe, B. Predicting mutational effects on protein binding from folding energy, (2025). URL http://arxiv.org/abs/2507.05502. arXiv:2507.05502 [q-bio].

Ding, D., Hill, F., Santoro, A., Reynolds, M., and Botvinick, M. Attention over learned object embeddings enables complex visual reasoning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9112–9124. Curran Associates, Inc., (2021). URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2021/file/4c26774d852f62440fc746ea4cdd57f6-Paper.pdf.

Feldman, J. and Feldman, T. Resilient biosecurity in the era of ai-enabled bioweapons, (2025). URL https://arxiv.org/abs/2509.02610.

Feldman, J. and Skolnick, J. AF3Complex Yields Improved Structural Predictions of Protein Complexes, (2025). URL https://www.biorxiv.org/content/10.1101/2025.02.27.640585v1. Pages: 2025.02.27.640585 Section: New Results.

Gao, J., Zheng, S., Yao, M., and Wu, P. (2021). Precise estimation of residue relative solvent accessible area from $c\alpha$ atom distance matrix using a deep learning method. *Bioinformatics*, 38(1):94–98. doi: 10.1093/bioinformatics/btab616.

Girdhar, R. and Ramanan, D. Attentional pooling for action recognition. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., (2017). URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2017/file/67c6a1e7ce56d3d6fa748ab6d9af3fd7-Paper.pdf.

Graber, D., Stockinger, P., Meyer, F., Mishra, S., Horn, C., and Buller, R. (2025). Resolving data bias improves generalization in binding affinity prediction. *Nature Machine Intelligence*, 7(10):1713–1725. doi: 10.1038/s42256-025-01124-5. Publisher: Nature Publishing Group.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., Badkundri, R., Shafkat, I., Gong, J., Derry, A., Molina, R. S., Thomas, N., Khan, Y., Mishra, C., Kim, C., Bartie, L. J., Nemeth, M., Hsu, P. D., Sercu, T., Candido, S., and Rives, A. Simulating 500 million years of evolution with a language model, (2024). URL https://www.biorxiv.org/content/10.1101/2024.07.01.600583v1. Pages: 2024.07.01.600583 Section: New Results.

Hummer, A. M., Schneider, C., Chinery, L., and Deane, C. M. (2025). Investigating the volume and diversity of data needed for generalizable antibody–antigen $\Delta\Delta G$ prediction. *Nature Computational Science*, pages 1–13. doi: 10.1038/s43588-025-00823-8. Publisher: Nature Publishing Group.

Huot, M., Wang, D., Liu, J., and Shakhnovich, E. I. (2025). Predicting high-fitness viral protein variants with bayesian active learning and biophysics. *Proceedings of the National Academy of Sciences*, 122(24):e2503742122.

Jankauskaitė, J., Jiménez-García, B., Dapkūnas, J., Fernández-Recio, J., and Moal, I. H. (2018). Skempi 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469. doi: 10.1093/bioinformatics/bty635.

Jiralerspong, M., Bose, J., Gemp, I., Qin, C., Bachrach, Y., and Gidel, G. Feature likelihood divergence: Evaluating the generalization of generative models using samples. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 33095–33119. Curran Associates, Inc., (2023). URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2023/file/68b138608ef80b08d65b1bd9594d9559-Paper-Conference.pdf.

Knudsen, M. and Wiuf, C. (2010). The *CATH* database. *Human Genomics*, 4(3):207–212. doi: 10.1186/1479-7364-4-3-207.

Koehl, P. and Levitt, M. (2002). Sequence variations within protein families are linearly related to structural variations. *Journal of Molecular Biology*, 323(3):551–562. doi: 10.1016/s0022-2836(02)00971-3.

Krissinel, E. (2007). On the relationship between sequence and structure similarities in proteomics. *Bioinformatics*, 23(6):717–723. doi: 10.1093/bioinformatics/btm006.

Lee, J., Lee, I., and Kang, J. Self-attention graph pooling. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3734–3743. PMLR, (2019). URL https://proceedings.mlr.press/v97/lee19c.html.

Li, G., Yao, S., and Fan, L. (2024). ProSTAGE: Predicting Effects of Mutations on Protein Stability by Using Protein Embeddings and Graph Convolutional Networks. *Journal of Chemical Information and Modeling*, 64(2):340–347. doi: 10.1021/acs.jcim.3c01697. Publisher: American Chemical Society.

Li, M., Tan, P., Ma, X., Zhong, B., Yu, H., Zhou, Z., Ouyang, W., Zhou, B., Hong, L., and Tan, Y. ProSST: Protein Language Modeling with Quantized Structure and Disentangled Attention, (2024). URL https://www.biorxiv.org/content/10.1101/2024.04.15.589672v3. Pages: 2024.04.15.589672 Section: New Results.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, 22(13):1658–1659. doi: 10.1093/bioinformatics/btl158.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130. doi: 10.1126/science.ade2574.

Loux, T., Wang, D., and Shakhnovich, E. I. (2024). More structure, less accuracy: Esm3's binding prediction paradox. *bioRxiv*, pages 2024–12.

Luo, S., Su, Y., Wu, Z., Su, C., Peng, J., and Ma, J. (2023). Rotamer density estimator is an unsupervised learner of the effect of mutations on protein-protein interaction. *bioRxiv*. doi: 10.1101/2023.02.28.530137.

Mariani, V., Biasini, M., Barbato, A., and Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728. doi: 10.1093/bioinformatics/btt473.

Marquet, C., Heinzinger, M., Olenyi, T., Dallago, C., Erckert, K., Bernhofer, M., Nechaev, D., and Rost, B. (2022). Embeddings from protein language models predict conservation and variant effects. *Human Genetics*, 141(10):1629–1647. doi: 10.1007/s00439-021-02411-y.

Mirabello, C. and Wallner, B. (2024). DockQ v2: improved automatic quality measure for protein multimers, nucleic acids, and small molecules. *Bioinformatics*, 40(10):btae586. doi: 10.1093/bioinformatics/btae586.

Mitternacht, S. (2016). Freesasa: An open source c library for solvent accessible surface area calculations. *F1000Research*, 5:189. doi: 10.12688/f1000research.7931.1.

Mo, Y., Hong, X., Gao, B., Jia, Y., and Lan, Y. Multi-level Interaction Modeling for Protein Mutational Effect Prediction, (2024). URL http://arxiv.org/abs/2405.17802. arXiv:2405.17802 [cs].

Park, H., Bradley, P., Greisen, P., Liu, Y., Mulligan, V. K., Kim, D. E., Baker, D., and DiMaio, F. (2016). Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *Journal of Chemical Theory and Computation*, 12(12):6201–6212. doi: 10.1021/acs.jctc.6b00819.

Pucci, F., Schwersensky, M., and Rooman, M. (2022). Artificial intelligence challenges for predicting the impact of mutations on protein stability. *Current Opinion in Structural Biology*, 72:161–168. doi: https://doi.org/10.1016/j.sbi.2021.11.001.

Roelofs, R., Shankar, V., Recht, B., Fridovich-Keil, S., Hardt, M., Miller, J., and Schmidt, L. A meta-analysis of overfitting in machine learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., (2019). URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2019/file/ee39e503b6bedf0c98c388b7e8589aca-Paper.pdf.

Sanavia, T., Birolo, G., Montanucci, L., Turina, P., Capriotti, E., and Fariselli, P. (2020). Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. *Computational and Structural Biotechnology Journal*, 18:1968–1979. doi: https://doi.org/10.1016/j.csbj.2020.07.011.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Research*, 33(Web Server issue): W382–388. doi: 10.1093/nar/gki387.

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H. D., Dingens, A. S., Navarro, M. J., Bowen, J. E., Tortorici, M. A., Walls, A. C., King, N. P., Veesler, D., and Bloom, J. D. (2020). Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*, 182(5):1295–1310.e20. doi: 10.1016/j.cell.2020.08.012.

Starr, T. N., Czudnochowski, N., Liu, Z., Zatta, F., Park, Y.-J., Addetia, A., Pinto, D., Beltramello, M., Hernandez, P., Greaney, A. J., et al. (2021). Sars-cov-2 rbd antibodies that maximize breadth and resistance to escape. *Nature*, 597(7874):97–102.

Tang, W., Zhang, W., and Zhang, M.-L. Disambiguated attention embedding for multi-instance partial-label learning. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 56756–56771. Curran Associates, Inc., (2023). URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2023/file/b1917a4bcfab403c3cdd6c6bbaf9fda0-Paper-Conference.pdf.

Thadani, N. N., Gurev, S., Notin, P., Youssef, N., Rollins, N. J., Ritter, D., Sander, C., Gal, Y., and Marks, D. S. (2023). Learning from prepandemic data to forecast viral escape. *Nature*, 622(7984):818–825. doi: 10.1038/s41586-023-06617-0. Publisher: Nature Publishing Group.

Thomas, L. B., Risheh, A., and Forouzesh, N. (2025). BPS2025 - Multi-objective loss function for free energy calculations. *Biophysical Journal*, 124(3):489a. doi: 10.1016/j.bpj.2024.11.2576. Publisher: Elsevier.

Vieira, L. C., Handojo, M. L., and Wilke, C. O. (2025). Scaling down for efficiency: Medium-sized protein language models perform well at transfer learning on realistic datasets. *bioRxiv*. doi: 10.1101/2024.11.22.624936.

Wang, D., Huot, M., Zhang, Z., Jiang, K., Shakhnovich, E. I., and Esvelt, K. M. Without safeguard, ai-bio integration risks accelerating future pandemics. *Frontiers in Microbiology*, 16:1734561.

Wang, D., Huot, M., Mohanty, V., and Shakhnovich, E. I. (2024). Biophysical principles predict fitness of sars-cov-2 variants. *Proceedings of the National Academy of Sciences*, 121(23):e2314518121. doi: 10.1073/pnas.2314518121.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., De Bortoli, V., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M., and Baker, D. (2023). De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100. doi: 10.1038/s41586-023-06415-8. Publisher: Nature Publishing Group.

Wee, J. and Wei, G.-W. (2024). Benchmarking AlphaFold3's protein-protein complex accuracy and machine learning prediction reliability for binding free energy changes upon mutation. *ArXiv*, page arXiv:2406.03979v1.

Xu, D., Ruan, C., Kumar, S., Korpeoglu, E., and Achan, K. Self-attention with functional time representation learning, (2019). URL https://arxiv.org/abs/1911.12864.

Yin, R. and Pierce, B. G. (2024). Evaluation of AlphaFold antibody–antigen modeling with implications for improving predictive accuracy. *Protein Science : A Publication of the Protein Society*, 33(1):e4865. doi: 10.1002/pro.4865.

Yu, G., Zhao, Q., Bi, X., and Wang, J. (2024). DDAffinity: predicting the changes in binding affinity of multiple point mutations using protein 3D structure. *Bioinformatics*, 40 (Supplement_1):i418–i427. doi: 10.1093/bioinformatics/btae232.

Yu, G., Bi, X., Ma, T., Li, Y., and Wang, J. (2025). CATH-ddG: towards robust mutation effect prediction on protein–protein interactions out of CATH homologous superfamily. *Bioinformatics*, 41(Supplement_1):i362–i372. doi: 10.1093/bioinformatics/btaf228.

Zhang, C., Shine, M., Pyle, A. M., and Zhang, Y. (2022). Us-align: Universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *bioRxiv*. doi: 10.1101/2022.04.18.488565.

Zhao, H., Velez, C., Navarene, A., Saha, A., Feldman, J., Skolnick, J., Murray, D., and Honig, B. (2025). Combining structural modeling and deep learning to calculate the e. coli protein interactome and functional networks. *bioRxiv*. doi: 10.1101/2025.05.07.652715.