Biophysical Priors Enhance Protein-Protein Binding $\Delta\Delta G$ Prediction

Antoine Maechler[1,2,⋆], **Jonathan Feldman**[1,3,⋆], **Dianzhuo Wang**[1,⋆,†], **Eugene I. Shakhnovich**[1,†]
[1]Department of Chemistry and Chemical Biology, Harvard University; Cambridge, MA 02138, USA
[2]Ecole Polytechnique, Institut Polytechnique de Paris, France
[3]College of Computing, Georgia Institute of Technology; Atlanta, GA, USA
⋆Equal contribution
†johnwang@g.harvard.edu, shakhnovich@chemistry.harvard.edu

Abstract

Predicting mutational effects on protein–protein binding affinity ($\Delta\Delta G$) remains a challenging task: current models often generalize poorly due to limited and biased training data. We show that SKEMPI2, the dominant training and evaluation dataset in this field, is affected by a subtle and pervasive data leakage due to sequential and structural redundancy, leading to inflated estimates of performance across models. We introduce ProtBFF (Protein Biophysical Feature Framework), a lightweight, encoder-agnostic module that injects five key biophysical features (interface, burial, dihedral, SASA, lDDT) into residue latent representations via cross-embedding attention. ProtBFF consistently improves predictive power and, with ProSST, achieves state-of-the-art performance on clustered SKEMPI2, rivaling far more specialized models. These results point to a simple, general recipe for protein property prediction: integrate biophysical priors with machine learning.

1 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 and therapeutic design [1–4]. Biophysics-based methods such as molecular dynamics have long been used for this task, but they are computationally expensive and rely on human-crafted priors, which limits their scalability[5–7].

While deep learning has transformed protein structure prediction [8–10], predicting $\Delta\Delta G$ remains constrained by small, biased datasets and models that often fail to generalize to previously unseen proteins[11–13]. This shortcoming arises from two main factors: (1) the restricted quality and quantity of available experimental measurements, compounded by the limited diversity of the proteins represented in existing datasets, and (2) the tendency of models to overfit dataset-specific patterns rather than capture underlying biophysical principles.

A critical example is the SKEMPI2 dataset, the field's foremost benchmark. While SKEMPI2 contains roughly 350 curated protein complexes, it suffers from substantial sequential and structural redundancy, with many complexes exhibiting high similarity [14, 15]. As we show and quantitatively evaluate in this work, such hidden redundancy introduces significant data leakage between training and test sets, inflating reported performance and obscuring the true generalization gap in current methods [14, 15].

To tackle this challenge, we leverage the complementary strengths of biophysics-based and deep learning methods: the mechanistic grounding of biophysical principles and the statistical power of large-scale representation learning. We introduce ProtBFF (Protein Biophysical Feature Framework), a lightweight, encoder-agnostic module designed to combine these advantages by enriching embedding-based predictors with explicit biophysical features. By scaling residue embeddings according to local

Machine Learning for Structural Biology Workshop

structural context and integrating them through cross-embedding attention, ProtBFF encourages models to learn representations that are both physically meaningful and predictive. Unlike prior work that proposes entirely new $\Delta\Delta G$ architectures [16–18] or relies solely on physics-based scoring functions [5, 7], **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 [19] and ESM family [9], to **reach or surpass the performance of state-of-the-art specialized PPI predictors.**

2 Limitations of the SKEMPI2 Dataset

2.1 Latent Sequential and Structural Data Leakage

The key problem of the SKEMPI2 dataset is a systematic data leakage. Prior studies split training and test sets by clustering mutations according to the complex name from which they were derived [16–18, 20]. Although intended to prevent overlap, this strategy ignores that many differently named complexes are highly homologous in both sequence and structure[14, 15, 21]. Consequently, homologous proteins often appear across both splits, inflating reported performance and introducing systematic data leakage.

To quantify this effect, we adopted a stricter dataset splitting strategy based on sequence similarity. Specifically, we used the CD-HIT clustering algorithm [22] to group protein complexes by sequence identity, applying thresholds from 60% to 100%. In this scheme, two complexes are assigned to the same cluster if the concatenated sequences of their chains exceed the chosen similarity threshold. At 100% identity, this procedure is equivalent to no clustering, producing one cluster per complex (335 in total; see Appendix A.1). At the other extreme, a 60% threshold reduces the dataset to only 136 clusters, underscoring the strong homology within SKEMPI2. Notably, even a modest reduction from 100% to 99% identity decreases the number of clusters from 335 to 253, indicating that many complexes differ by only minimal sequence variation. This sharp drop highlights the extent of redundancy in SKEMPI2 and demonstrates why clustering by complex name alone is insufficient to prevent overlap. Table 1 summarizes the number of clusters across thresholds.

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.

2.2 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 [17], ProMIM [18], RDE-Network [16], and ProSST [19], using 10-fold cross-validation on these clustered datasets. At lower sequence-identity thresholds, training and test sets contain fewer homologous proteins, so performance should decline if earlier results were inflated by redundancy. This setup provides a systematic test of model robustness.

As shown in Figure 1, while models achieve high Pearson correlations of above 0.6 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. Details about the methodology and analogous results using Spearman's correlation are provided in Appendix Figure 3.

In summary, these results demonstrate that deep learning models trained on SKEMPI2 are subject to substantial overfitting due to data leakage, and that previously reported performance metrics are overly optimistic. This underscores the need for more robust benchmarks and generalizable architectures, which we address in the remainder of this paper.

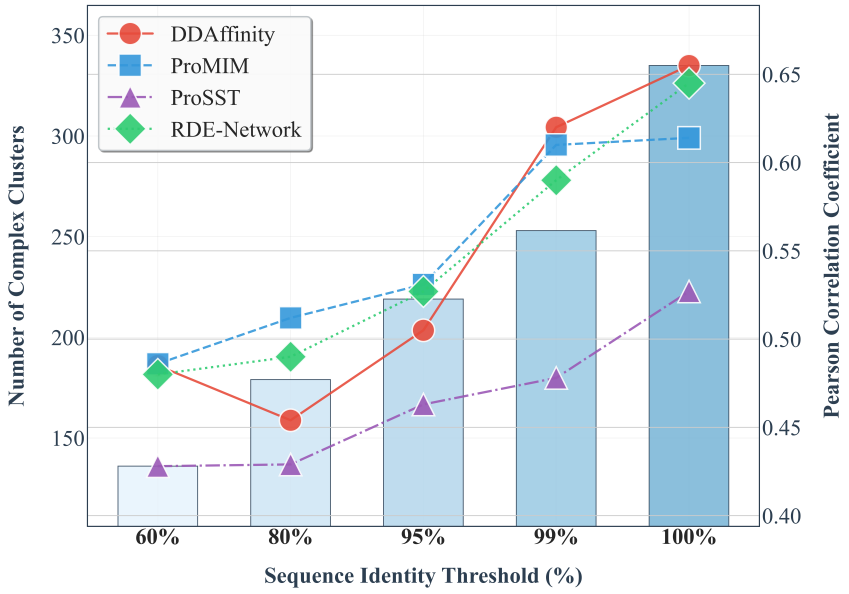

Figure 1: Performance of deep learning models on SKEMPI2 under different sequence identity thresholds. Bars indicate the number of complex clusters at each threshold, and lines show Pearson correlation for each model. Performance drops sharply as clustering becomes more stringent, revealing substantial data leakage in the original dataset splits.

3 Introducing ProtBFF

3.1 Methods: 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 $\Delta\Delta G$ [23–25]. Concretely, an encoder generates per-residue embeddings for both wildtype and mutant proteins, integrating sequential and structural information [18, 16, 17]. These encoders are typically pretrained on large-scale datasets with self-supervised objectives (e.g., masked residue prediction) to build informative internal representations [23, 19]. For $\Delta\Delta G$ prediction, embeddings of the wildtype and mutant are subtracted and then pooled across residues, and the resulting features are passed to downstream predictors trained on task-specific datasets such as SKEMPI2 [16, 17, 26]. In this study, we propose a general biophysical framework that augments such models by integrating explicit biophysical features into per-residue embeddings (Figure 2).

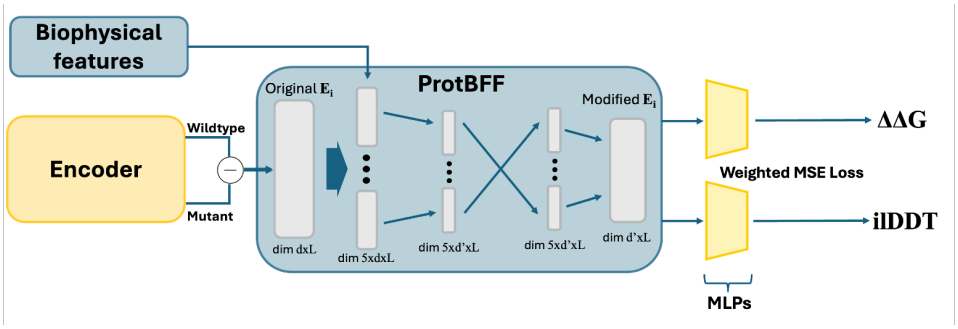

Figure 2: Simplified schematic of the ProtBFF framework. The pipeline begins with embeddings extraction, followed by embeddings subtraction and scaling using biophysical features. These processed embeddings are passed through a pooling layer, followed by the cross-embedding attention module and two MLP heads to generate predictions of $\Delta\Delta G$ and $ilDDT$, optimized with a weighted loss function. For more information, refer to Appendix B.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 [17, 27]. 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 [28, 29]. 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). These metrics are computed from the wildtype structures and FoldX-generated mutant structures [5] (see Appendix D).

Formally, for residue i with embedding $\mathbf{E}_i \in \mathbb{R}^M$, 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 tasks[30].

As illustrated in Fig. 2, 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.[30, 31] Finally, a lightweight 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 Appendix B.1 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 ilDDT [32, 33]. This auxiliary prediction encourages the extraction of structurally meaningful features and improves generalization. More information can be found in Appendix B.2

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.

3.2 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 [19], and the ESM2 [34] and ESM3 models [9], 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 Appendix C.1.

ProSST [19], though not originally designed for protein complex $\Delta\Delta G$ prediction, gains substantially with ProtBFF: Pearson improves from 0.428 $\rightarrow$ 0.514 and Spearman from 0.354 $\rightarrow$ 0.477, surpassing specialized models such as ProMIM and DDAffinity. ESM models also benefit substantially, with Pearson and Spearman correlations increasing from below 0.2 to levels comparable to most state-of-the-art models, thereby more than doubling their baseline performance.

These results illustrate that models pretrained on large protein datasets, such as CATH database [35] for ProSST and the extensive sequence and structural corpora underlying ESM2 and ESM3, learn rich general-purpose representations [9, 19]. 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 Appendices B.2, and C.

Benchmarking and Ablation Results	Baseline		With ProtBFF	
Method / Variant	Pearson (ρ)	Spearman (r)	Pearson (ρ)	Spearman (r)
Model Benchmarking				
ProSST [19]	0.428	0.354	**0.514**	**0.477**
ESM2 [34]	0.194	0.204	0.451	0.410
ESM3 [9]	0.159	0.099	0.362	0.347
ProMIM [18]	0.486	0.464	*	*
RDE-Network [16]	0.480	0.439	*	*
DDAffinity [17]	0.485	0.405	*	*
RDE-Linear [16]	0.369	0.360	*	*
FoldX [5]	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 correlation (r) and Spearman correlation (ρ). Bold indicates the best performance. Asterisks (*) mark methods where ProtBFF cannot be applied because they are not encoder-based. See Appendix C for details.

Ablation Study: To assess the contribution of the five biophysical scores and the auxiliary ilDDT loss to ProtBFF's performance, we conducted an ablation study (Appendix E). Each component was systematically removed in turn, and we observed that all scores as well as the ilDDT loss provided measurable improvements in both Pearson and Spearman correlations. Among them, 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.

4 Discussion

The challenge of predicting $\Delta\Delta G$ highlights a core tension in machine learning: building models that generalize when data is scarce and biased. While self-supervised pretraining on massive sequence databases has transformed protein structure prediction [8–10], supervised $\Delta\Delta G$ prediction remains limited by the small size of datasets like SKEMPI2, which contains only 7,085 measurements across 345 complexes [2, 14, 15, 36].

Our analysis shows that splitting SKEMPI2 by complex name overlooks deep homology, leading to data leakage and inflated accuracy. When clustered by sequence identity, model performance drops sharply, revealing that current methods often memorize dataset-specific patterns rather than learn transferable biophysics [37, 38], echoing test-set contamination issues in other ML fields [39].

ProtBFF addresses this gap by injecting biophysical context into embeddings. Its five feature-based masks emphasize structurally relevant residues, while cross-embedding attention integrates these signals, yielding representations that generalize better. Notably, ProtBFF boosts models like ProSST and ESM—encoders not designed for binding prediction—demonstrating its value as a lightweight adapter for transfer learning.

Looking ahead, progress will require benchmarks with cluster-based splits that reflect evolutionary diversity and models that combine learned representations with physics-based features across scales. ProtBFF illustrates that robust protein engineering tools arise not from choosing between physics and ML, but from integrating them.

References

[1] Tiziana Sanavia, Giovanni Birolo, Ludovica Montanucci, Paola Turina, Emidio Capriotti, and Piero Fariselli. Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. *Computational and Structural Biotechnology Journal*, 18:1968–1979, 2020. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2020.07.011. URL https://www.sciencedirect.com/science/article/pii/S2001037020303433.

[2] Fabrizio Pucci, Martin Schwersensky, and Marianne Rooman. Artificial intelligence challenges for predicting the impact of mutations on protein stability. *Current Opinion in Structural Biology*, 72:161–168, 2022. ISSN 0959-440X. doi: https://doi.org/10.1016/j.sbi.2021.11.001. URL https://www.sciencedirect.com/science/article/pii/S0959440X21001445.

[3] Marian Huot, Dianzhuo Wang, Jiacheng Liu, and Eugene I Shakhnovich. Predicting high-fitness viral protein variants with bayesian active learning and biophysics. *Proceedings of the National Academy of Sciences*, 122(24):e2503742122, 2025.

[4] Dianzhuo Wang, Marian Huot, Vaibhav Mohanty, and Eugene I. Shakhnovich. Biophysical principles predict fitness of sars-cov-2 variants. *Proceedings of the National Academy of Sciences*, 121(23):e2314518121, 2024. doi: 10.1073/pnas.2314518121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2314518121.

[5] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The FoldX web server: an online force field. *Nucleic Acids Research*, 33(Web Server issue):W382–388, July 2005. ISSN 1362-4962. doi: 10.1093/nar/gki387.

[6] Hahnbeom Park, Philip Bradley, Per Greisen, Yuan Liu, Vikram Khipple Mulligan, David E. Kim, David Baker, and Frank DiMaio. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *Journal of Chemical Theory and Computation*, 12(12):6201–6212, December 2016. ISSN 1549-9626. doi: 10.1021/acs.jctc.6b00819.

[7] Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Jr. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation*, 13(6):3031–3048, June 2017. ISSN 1549-9618. doi: 10.1021/acs.jctc.7b00125. URL https://doi.org/10.1021/acs.jctc.7b00125. Publisher: American Chemical Society.

[8] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630 (8016):493–500, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL https://www.nature.com/articles/s41586-024-07487-w. Publisher: Nature Publishing Group.

[9] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model, July 2024. URL https://www.biorxiv.org/content/10.1101/2024.07.01.600583v1. Pages: 2024.07.01.600583 Section: New Results.

[10] Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL https://www.nature.com/articles/s41586-023-06415-8. Publisher: Nature Publishing Group.

[11] Alissa M. Hummer, Constantin Schneider, Lewis Chinery, and Charlotte M. Deane. Investigating the volume and diversity of data needed for generalizable antibody–antigen $\Delta\Delta$G prediction. *Nature Computational Science*, pages 1–13, July 2025. ISSN 2662-8457. doi: 10.1038/s43588-025-00823-8. URL https://www.nature.com/articles/s43588-025-00823-8. Publisher: Nature Publishing Group.

[12] Cunliang Geng, Li C. Xue, Jorge Roel-Touris, and Alexandre M. J. J. Bonvin. Finding the $\Delta\Delta$g spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? *WIREs Computational Molecular Science*, 9(5):e1410, 2019. doi: https://doi.org/10.1002/wcms.1410. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1410.

[13] Thomas Loux, Dianzhuo Wang, and Eugene I Shakhnovich. More structure, less accuracy: Esm3's binding prediction paradox. *bioRxiv*, pages 2024–12, 2024.

[14] Anton Bushuiev, Roman Bushuiev, Petr Kouba, Anatolii Filkin, Marketa Gabrielova, Michal Gabriel, Jiri Sedlar, Tomas Pluskal, Jiri Damborsky, Stanislav Mazurenko, and Josef Sivic. Learning to design protein-protein interactions with enhanced generalization, 2024. URL https://arxiv.org/abs/2310.18515.

[15] Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. Skempi 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 07 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty635. URL https://doi.org/10.1093/bioinformatics/bty635.

[16] Shitong Luo, Yufeng Su, Zuofan Wu, Chenpeng Su, Jian Peng, and Jianzhu Ma. Rotamer density estimator is an unsupervised learner of the effect of mutations on protein-protein interaction. *bioRxiv*, 2023. doi: 10.1101/2023.02.28.530137. URL https://www.biorxiv.org/content/early/2023/03/01/2023.02.28.530137.

[17] Guanglei Yu, Qichang Zhao, Xuehua Bi, and Jianxin Wang. DDAffinity: predicting the changes in binding affinity of multiple point mutations using protein 3D structure. *Bioinformatics*, 40 (Supplement_1):i418–i427, July 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae232. URL https://doi.org/10.1093/bioinformatics/btae232.

[18] Yuanle Mo, Xin Hong, Bowen Gao, Yinjun Jia, and Yanyan Lan. Multi-level Interaction Modeling for Protein Mutational Effect Prediction, May 2024. URL http://arxiv.org/abs/2405.17802. arXiv:2405.17802 [cs].

[19] Mingchen Li, Pan Tan, Xinzhu Ma, Bozitao Zhong, Huiqun Yu, Ziyi Zhou, Wanli Ouyang, Bingxin Zhou, Liang Hong, and Yang Tan. ProSST: Protein Language Modeling with Quantized Structure and Disentangled Attention, May 2024. URL https://www.biorxiv.org/content/10.1101/2024.04.15.589672v3. Pages: 2024.04.15.589672 Section: New Results.

[20] JunJie Wee and Guo-Wei Wei. Evaluation of AlphaFold 3's Protein–Protein Complexes for Predicting Binding Free Energy Changes upon Mutation. *Journal of Chemical Information and Modeling*, 64(16):6676–6683, August 2024. ISSN 1549-9596. doi: 10.1021/acs.jcim.4c00976. URL https://doi.org/10.1021/acs.jcim.4c00976. Publisher: American Chemical Society.

[21] Guanglei Yu, Xuehua Bi, Teng Ma, Yaohang Li, and Jianxin Wang. CATH-ddG: towards robust mutation effect prediction on protein–protein interactions out of CATH homologous superfamily. *Bioinformatics*, 41(Supplement_1):i362–i372, July 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf228. URL https://doi.org/10.1093/bioinformatics/btaf228.

[22] Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, 22(13):1658–1659, July 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl158.

[23] Céline Marquet, Michael Heinzinger, Tobias Olenyi, Christian Dallago, Kyra Erckert, Michael Bernhofer, Dmitrii Nechaev, and Burkhard Rost. Embeddings from protein language models predict conservation and variant effects. *Human Genetics*, 141(10):1629–1647, 2022. ISSN 0340-6717. doi: 10.1007/s00439-021-02411-y. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8716573/.

[24] Gen Li, Sijie Yao, and Long Fan. 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, January 2024. ISSN 1549-9596. doi: 10.1021/acs.jcim.3c01697. URL https://doi.org/10.1021/acs.jcim.3c01697. Publisher: American Chemical Society.

[25] Ashim Dahal, Saydul Akbar Murad, and Nick Rahimi. Embedding Shift Dissection on CLIP: Effects of Augmentations on VLM's Representation Learning, April 2025. URL http://arxiv.org/abs/2503.23495. arXiv:2503.23495 [cs].

[26] Shiwei Liu, Tian Zhu, Milong Ren, Chungong Yu, Dongbo Bu, and Haicang Zhang. Predicting mutational effects on protein-protein binding via a side-chain diffusion probabilistic model, 2023. URL https://arxiv.org/abs/2310.19849.

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

[28] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 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.

[29] Wei Tang, Weijia Zhang, and Min-Ling Zhang. Disambiguated attention embedding for multi-instance partial-label learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 56756–56771. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/b1917a4bcfab403c3cdd6c6bbaf9fda0-Paper-Conference.pdf.

[30] David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matt Botvinick. Attention over learned object embeddings enables complex visual reasoning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, 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.

[31] Da Xu, Chuanwei Ruan, Sushant Kumar, Evren Korpeoglu, and Kannan Achan. Self-attention with functional time representation learning, 2019. URL https://arxiv.org/abs/1911.12864.

[32] Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, November 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt473. URL https://doi.org/10.1093/bioinformatics/btt473.

[33] Libina Bovan Thomas, Ali Risheh, and Negin Forouzesh. BPS2025 - Multi-objective loss function for free energy calculations. *Biophysical Journal*, 124(3):489a, February 2025. ISSN 0006-3495, 1542-0086. doi: 10.1016/j.bpj.2024.11.2576. URL https://www.cell.com/biophysj/abstract/S0006-3495(24)03304-6. Publisher: Elsevier.

[34] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

[35] M. Knudsen and C. Wiuf. The *CATH* database. *Human Genomics*, 4(3):207–212, Feb 2010. doi: 10.1186/1479-7364-4-3-207. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3525972/.

[36] Arthur Deng, Karsten Householder, Fang Wu, Sebastian Thrun, K. Christopher Garcia, and Brian Trippe. Predicting mutational effects on protein binding from folding energy, July 2025. URL http://arxiv.org/abs/2507.05502. arXiv:2507.05502 [q-bio].

[37] Marco Jiralerspong, Joey Bose, Ian Gemp, Chongli Qin, Yoram Bachrach, and Gauthier Gidel. Feature likelihood divergence: Evaluating the generalization of generative models using samples. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 33095–33119. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/68b138608ef80b08d65b1bd9594d9559-Paper-Conference.pdf.

[38] Xingyi Cheng, Bo Chen, Pan Li, Jing Gong, Jie Tang, and Le Song. Training compute-optimal protein language models, 2024. URL https://arxiv.org/abs/2411.02142.

[39] Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. A meta-analysis of overfitting in machine learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/ee39e503b6bedf0c98c388b7e8589aca-Paper.pdf.

[40] Evgeny Krissinel. On the relationship between sequence and structure similarities in proteomics. *Bioinformatics*, 23(6):717–723, March 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm006. URL https://doi.org/10.1093/bioinformatics/btm006.

[41] Patrice Koehl and Michael Levitt. Sequence variations within protein families are linearly related to structural variations. *Journal of Molecular Biology*, 323(3):551–562, October 2002. ISSN 0022-2836. doi: 10.1016/s0022-2836(02)00971-3.

[42] M. Biasini, T. Schmidt, S. Bienert, V. Mariani, G. Studer, J. Haas, N. Johner, A. D. Schenk, A. Philippsen, and T. Schwede. *OpenStructure*: an integrated software framework for computational structural biology. *Acta Crystallographica Section D*, 69(5):701–709, May 2013. doi: 10.1107/S0907444913007051. URL https://doi.org/10.1107/S0907444913007051.

[43] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3734–3743. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/lee19c.html.

[44] Wataru Nishima, Guoying Qi, Steven Hayward, and Akio Kitao. Dta: dihedral transition analysis for characterization of the effects of large main-chain dihedral changes in proteins. *Bioinformatics*, 25(5):628–635, 01 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp032. URL https://doi.org/10.1093/bioinformatics/btp032.

[45] Jianzhao Gao, Shuangjia Zheng, Mengting Yao, and Peikun Wu. Precise estimation of residue relative solvent accessible area from cα atom distance matrix using a deep learning method. *Bioinformatics*, 38(1):94–98, 08 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab616. URL https://doi.org/10.1093/bioinformatics/btab616.

[46] Simon Mitternacht. Freesasa: An open source c library for solvent accessible surface area calculations. *F1000Research*, 5:189, February 2016. ISSN 2046-1402. doi: 10.12688/f1000research.7931.1. URL http://dx.doi.org/10.12688/f1000research.7931.1.

A Dataset Collation

A.1 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) [15]. 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 [19]. 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 [22]. This step mitigated the data leakage issue in SKEMPI, since high sequence similarity often implies structural similarity[40, 41]. Given that SKEMPI2 contains many complexes with nearly identical composition, this clustering also reduced structural homology overlap across folds [15].

A.2 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 [5]. Importantly, this procedure does not modify the protein backbone. Although not explored here, alternative methods that incorporate backbone relaxation, such as Rosetta, 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 [7]. Among the five ProtBFF features, only dihedral and lDDT require FoldX mutant structures, since they depend on comparing WT and mutant conformations. The other three features:interface, burial, and SASA are computed solely from the WT structure and do not rely on mutant modeling.

B Further Framework Details

B.1 Embedding-aware attention Network Architecture

The ProtBFF framework operates by processing 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. In the standard case, L corresponds to the full sequence length of the complex; however, L may also represent a subset of residues identified as important by the encoder. In such cases, a mapping relation is used to relate positions in the full sequence to those retained in the reduced set, and scaling is applied only to those residues.

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 2

B.2 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 [32, 42]. 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:

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

where $\widehat{\Delta\Delta G}$ and $\widehat{\text{ilDDT}}$ 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 [28, 43]. During inference, the ilDDT output is ignored, as it serves only as a conditioning signal during training.

C $\Delta\Delta G$ 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 [9], while for ESM2, only sequences were embedded [34]. 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 [18, 16, 17].

C.1 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.

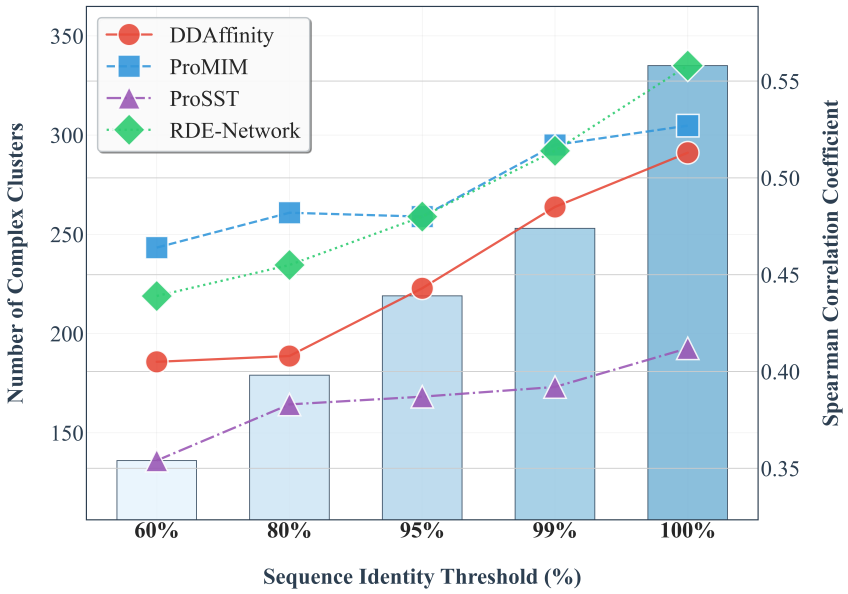

Figure 3: Performance of all predictive models across sequence identity thresholds in SKEMPI2 measured by Spearman rank correlation coefficient. Each line represents a different model with the legend identifying each model.

D 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.

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 [17]. 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 the structural changes in side-chain dihedral (chi) angles that occur after a mutation. The intuition behind this score is that mutations which cause significant alterations

in chi angles can lead to substantial conformational rearrangements, potentially impacting the overall stability and function of the protein. A higher dihedral score corresponds to a larger change in chi angles upon mutation, indicating a more pronounced structural shift [44]. The dihedral score for residue i is

$$\Delta\Phi_i = \frac{1}{360} \sum_{j \in N(i,9)} |\Delta\phi_j| \,,$$

where $N(i, 9)$ are its 9 nearest neighbors and $\Delta\phi_j$ is the absolute change in chi angle(s) between the wildtype and FoldX-generated mutant structures. Dividing by 360 normalizes the score to the range between 0 and 1.

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 [45]. This score is calculated using the *freesasa* package, which determines the SASA of each residue based on atomic coordinates provided in the wildtype PDB structure [46]. The SASA score incorporates both the solvent exposure of the residue of interest and the exposure of its nearby residues, weighted by their proximity.

The SASA score for residue i is the Gaussian-weighted sum of solvent-accessible surface areas for its 10 nearest 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} is a Gaussian weight favoring closer neighbors, $N(i, 10)$ is the 10 nearest neighbors of i, and $\{L\}$ is the set of all residues.

The ***lDDT score*** measures per-residue atomic conformational changes between the wildtype (template) and FoldX-generated mutant (predicted) structures [32]. Larger changes indicate residues more likely to affect overall structure and function. The score is normalized to $[0, 1]$, and we scale embeddings by $1 -$ lDDT so that residues with greater changes receive higher emphasis during max pooling.

E 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 feature removed at a time, while keeping the remaining components intact. This allowed for the quantification of their individual importance in driving $\Delta\Delta G$ prediction performance.

E.1 Experimental Setup

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 C.1.

E.2 Ablation Study Results

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, but still meaningful, reductions in performance, indicating that they provide complementary structural information that refines the model's predictions. Additionally, we tested the impact of removing the ilDDT-based loss function. While its effect is less pronounced than the biophysical scores themselves, we observed a measurable decline in both correlation metrics, verifying that incorporating structural fidelity into the training objective improves model robustness. Finally, a baseline variant without any biophysical scores performs

significantly worse, demonstrating that the multi-feature integration strategy is crucial for ProtBFF's predictive power.

E.3 Discussion of Ablation Results

These results confirm that each biophysical feature contributes uniquely to ProtBFF's predictive performance. The interface and residue scores encode the most directly relevant signals for mutational effects, while the SASA, dihedral, and lDDT scores provide complementary structural context that refines predictions. The ilDDT loss function further improves model robustness by encouraging structural fidelity during training, albeit with a smaller impact than the feature scores themselves. Importantly, no single feature or loss term is sufficient on its own, underscoring the necessity of integrating multiple biophysical signals alongside a structure-aware loss to achieve state-of-the-art $\Delta\Delta G$ prediction.