PROTEIN STRUCTURE PREDICTORS IMPLICITLY DEFINE BINDING ENERGY FUNCTIONS

Divya Nori, Anisha Parsan, Caroline Uhler
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA, 02139, USA
{divnor80,aparsan,cuhler}@mit.edu

Wengong Jin
Khoury College of Computer Science
Northeastern University
Boston, MA, 02115, USA
{w.jin}@northeastern.edu

ABSTRACT

Estimating binding energies is vital for drug discovery, yet supervised methods are hampered by limited experimental data. Recent protein structure predictors (e.g. AlphaFold3) offer unsupervised alternatives via confidence metrics that correlate with binding energies. However, these metrics operate on a fixed scale, limiting their ability to capture fine-grained energy differences. Leveraging the Joint Energy-based Model (JEM) framework, we show that protein structure predictors implicitly define an energy function, and we introduce two new energy-based models derived from the confidence head. Our EBMs consistently improve binding energy prediction, outperforming both traditional confidence metrics and unsupervised baselines, and demonstrate that structure prediction models can be repurposed as powerful unsupervised energy predictors.

1 INTRODUCTION

Accurate binding energy estimation is key to molecular engineering tasks like virtual screening and protein-protein interaction design. One approach is to apply supervised methods that predict binding free energies (ΔG) or their changes ($\Delta\Delta G$), but these rely on limited, costly experimental datasets. On the other hand, empirical energy functions (e.g., molecular mechanics or Rosetta scoring) are computationally intensive despite offering physical insight (Miller III et al., 2012; Schymkowitz et al., 2005). Recent approaches use confidence estimates from protein structure prediction models to address these challenges. For instance, interface pTM (ipTM) (Evans et al., 2021) correlates with experimental binding affinities and serves as an unsupervised proxy (Zambaldi et al., 2024). However, these confidence metrics operate on fixed, discretized scales and lack probabilistic interpretation, limiting their effectiveness for capturing continuous, fine-grained energy landscapes.

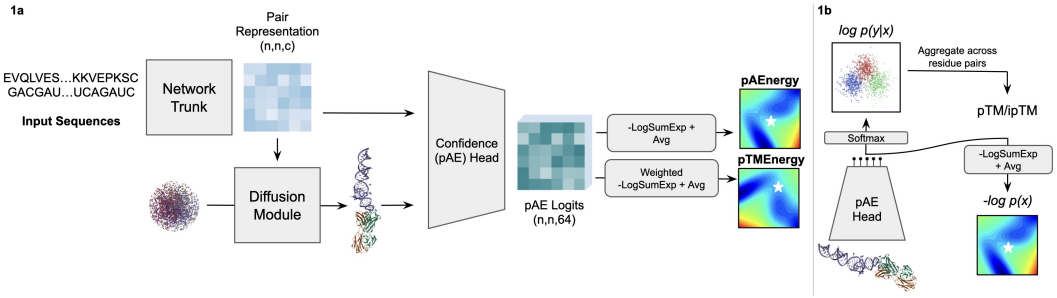

Figure 1: (a) Energy functions are derived from classifier outputs of protein structure predictors. (b) Application of JEM to protein structure predictors.

In this work, we extract energy functions directly from black-box structure prediction models using the Joint Energy-based Model (JEM) framework (Grathwohl et al., 2019). We derive two EBMs for scoring molecular interactions: pAEnergy, from the outputs of the confidence head, and pTMEnergy,

a weighted variant of pAEnergy. Unlike models that require extensive experimental data or high-resolution crystal structures, our method leverages pretrained predictors to capture rich interaction information. This yields a continuous energy landscape that overcomes the limitations of fixed-scale confidence scores.

2 Related Work

Please see Appendix A.1.

3 Methods: From confidence to energy

Motivation. State-of-the-art protein folding models provide confidence scores that have been used as binding affinity proxies in virtual screening (Bennett et al., 2023). However, these metrics operate on a binary scale, limiting their utility for tasks like $\Delta\Delta G$ prediction, which requires a continuous measure. For example, the Alphafold3 ipTM score difference shows poor correlation with true $\Delta\Delta G$ on the SKEMPI dataset (Pearson R = 0.102; Figure 2).

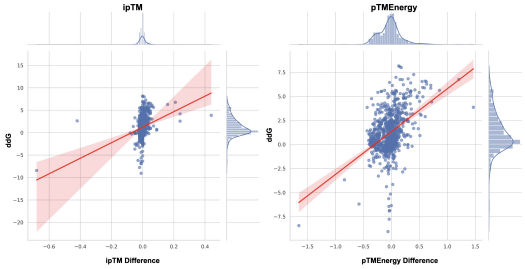

Figure 2: Scatter plots comparing ipTM and pTMEnergy by their association with $\Delta\Delta G$.

We propose an alternative measure of binding affinity by taking an energy-based modeling (EBM) view of protein folding models. Specifically, our goal is to extract an energy function $E_\theta(x)$ from any black-box protein structure predictor where x is the sequence or structure of a protein complex, depending on the predictor's input. Under the EBM framework, the probability density $p(x)$ is expressed as:

$$p_\theta(x) = \frac{\exp(-E_\theta(x))}{Z(\theta)} \tag{1}$$

where $E_\theta(x)$ represents the un-normalized log-likelihood (energy) of any input x. In the context of proteins, $E_\theta(x)$ can be viewed as a measure of binding affinity, where a lower energy means better stability. Compared to standard confidence metrics, this energy score gives a more continuous measure of binding affinity, which is more suitable for fine-grained tasks like $\Delta\Delta G$ prediction.

3.1 Casting protein folding models as EBMs

To obtain an EBM from a protein folding model, we leverage the fact that any classifier implicitly defines an energy function (Grathwohl et al., 2019). Consider a classification model f_θ which maps protein x to K real-valued logits. In the final softmax layer, these logits parameterize a categorical distribution where $f_\theta(x)[y]$ denotes the logit corresponding to the yth class label:

$$p_\theta(y|x) = \frac{\exp(f_\theta(x)[y])}{\sum_{y'} \exp(f_\theta(x)[y'])} \tag{2}$$

The logits can be reinterpreted to define an EBM over protein sequences or structures x and labels y:

$$p_\theta(x, y) = \frac{\exp(f_\theta(x)[y])}{Z(\theta)}, \tag{3}$$

where $E_\theta(x, y) = -f_\theta(x)[y]$. We can marginalize out y to arrive at a probability density for x and subsequently define an energy function $E_\theta(x)$:

$$p_\theta(x) = \frac{\sum_y \exp(f_\theta(x)[y])}{Z(\theta)} \tag{4}$$

$$E_\theta(x) = -\text{LogSumExp}_y(f_\theta(x)[y]) \tag{5}$$

Thus, we can leverage any classifier as an EBM, where $\text{LogSumExp}(\cdot)$ over the logits defines an energy function. To apply this method to protein structure predictors, we notice that most of them employ classifiers for the task of confidence (pAE/pTM) prediction. Based on this observation, we propose two EBMs for unsupervised binding energy prediction: pAEnergy and pTMEnergy. Since the confidence head's input is a predicted structure, these EBMs estimate energy of a predicted complex structure.

3.1.1 pAENERGY

Protein structure predictors employ a classification head in their confidence model which predicts the predicted aligned error (pAE) between residue pairs. For an input protein complex, consider the ground-truth structure's C_α atoms $X^{\text{true}} = \{\vec{x}_j^{\text{true}}\}$, its backbone frames $\{T_i^{\text{true}}\}$, predicted structure's C_α atoms $X = \{\vec{x}_j\}$, and corresponding backbone frames $\{T_i\}$. We can compute a pairwise error matrix $e_{ij} = T_i^{-1} \cdot \vec{x}_j - T_i^{\text{true}^{-1}} \cdot x_j^{\text{true}}$, capturing the error between true and predicted C_α positions when aligned on backbone frames. The magnitude of this error is then discretized into error bins to form a classification target.

Within the confidence model, the pAE head is trained to predict the discretized pAE matrix where $pAE[i, j, y]$ gives the probability that the alignment error magnitude between residues i and j falls into bin y. The logits are correspondingly denoted as $f_{\theta_{\text{pAE}}}(x)[i, j, y]$ and are supervised with a cross-entropy loss against the true alignment error. We define **pAEnergy** as an energy function derived from these logits:

$$E_{\text{pAE}}(x) = -\frac{1}{M} \sum_{i<j} \text{LogSumExp}_y(f_{\text{pAE}}(x)[i, j, y]) \tag{6}$$

3.1.2 pTMENERGY

As introduced by the original AlphaFold work (Jumper et al., 2021), the confidence model's predicted pAE matrix can be used to define a global confidence metric called the pTM score:

$$\text{pTM} = \max_i \frac{1}{N_{\text{res}}} \sum_j \mathbb{E}\big[g(e_{ij})\big] \tag{7}$$

The expectation is taken over the probability distribution over pairwise alignment errors bins between residues i and j, defined by e_{ij}. The scaling function g is defined as:

$$g(d_{ij}) = \frac{1}{1 + \left(\frac{d_{ij}}{d_0(N_{\text{res}})}\right)^2} \tag{8}$$

$$d_0(N_{\text{res}}) = 1.24 \sqrt[3]{\text{maximum}(N_{\text{res}}, 19) - 15} - 1.8 \tag{9}$$

The behavior of the pTM score is strongly influenced by g which weights each predicted bin probability within the expectation. Because $g(d_{ij})$ decreases as the predicted error bin grows, larger error bins receive a lower weight within the expectation. This incorporates the physical inductive bias that in real proteins, accurate local packing is critical for stability and function. We incorporate scaling function g and call this weighted energy function **pTMEnergy**:

$$E_{\text{pTM}}(x) = -\frac{1}{M} \sum_{i<j} E_{\text{pTM}}[i, j] \tag{10}$$

$$E_{\text{pTM}}[i, j] = \text{LogSumExp}_y\big(\log g(y) + f_{\text{pAE}}(x)[i, j, y]\big)$$

Following the approach of the ipTM score in AlphaFold-Multimer (Evans et al., 2021), we restrict the summation to interface residue pairs when computing all versions of our energy functions.

4 EXPERIMENTS

4.1 ENERGY PREDICTS PROTEIN-PROTEIN INTERACTION MUTATION EFFECTS

Protein-protein interactions (PPIs) are vital for many biological processes, and mutations at their interfaces can dramatically affect function (Cheng et al., 2021). Predicting the effect of a mutation is therefore valuable, quantified as the change in binding free energy $\Delta\Delta G = \Delta G^{mut} - \Delta G^{wt}$.

Experimental setup. We use SKEMPI v2 (Jankauskaitė et al., 2019), which provides $\Delta\Delta G$ measurements for 7,085 mutations across 348 protein complexes. Predicted $\Delta\Delta G$ is the difference between the predicted energies of mutant and wild-type complexes. We evaluate our method using AlphaFold3 (Abramson et al., 2024) and Chai-1 (Discovery et al., 2024). AlphaFold3 employs MSAs via Jackhmmer/Nhmmer (10 minutes per sample), whereas Chai-1 uses language model embeddings to bypass MSA computation. We compare against unsupervised $\Delta\Delta G$ rediction methods that do not require high-resolution crystal structures, including ESM-1v (Meier et al., 2021), PSSM, MSA Transformer (Rao et al., 2021), and Tranception (Notin et al., 2022). Additionally, we compare against DSMBind (Jin et al., 2023) and BA-Cycle (Jiao et al., 2024), state-of-the-art structure-based unsupervised binding prediction models. To ensure a fair comparison with our method which does not require crystal structures, we use AlphaFold3-predicted structures as the input to these models.

		Pearson	Spearman
Baselines	ESM-1v	0.192	0.157
	PSSM	0.016	0.067
	MSA Trans.	0.117	0.131
	Tranception	0.114	0.140
	DSMBind	0.101	0.067
	BA-Cycle	0.234	0.165
Chai-1	ipTM	$.142_{.007}$	$.158_{.009}$
	pAEnergy	$.188_{.004}$	$.139_{.013}$
	pTMEnergy	$.195_{.004}$	$.173_{.010}$
	Ensemble	$\mathbf{.268_{.006}}$	$\mathbf{.225_{.007}}$
AF3	ipTM	$.102_{.006}$	$.233_{.011}$
	pAEnergy	$.105_{.009}$	$.123_{.005}$
	pTMEnergy	$.265_{.003}$	$.208_{.004}$
	Ensemble	$\mathbf{.309_{.004}}$	$\mathbf{.278_{.007}}$

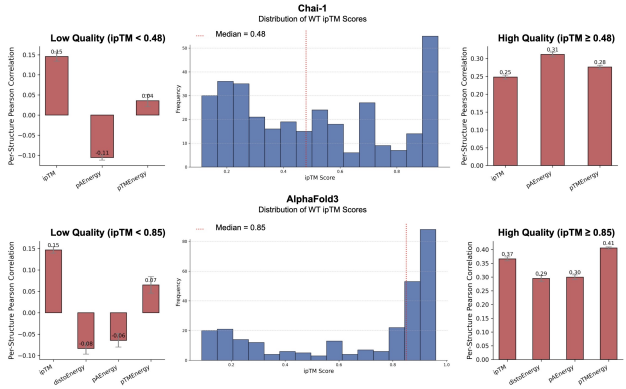

Figure 3: To the left, the table presents the overall correlation results on the SKEMPI v2 dataset. Chai-1 and AlphaFold3 results are averaged across five structure predictions per sequence. To the right, we illustrate the per-structure Pearson correlation.

Benchmark Results. pTMEnergy consistently outperforms baseline methods and standard confidence scores. For Chai-1, pTMEnergy improves Pearson correlation by 37% and Spearman correlation by nearly 10% compared to ipTM. To further improve prediction accuracy, we introduce an ensemble, which combines ipTM and pTMEnergy ranks with equal weight. This ensemble strategy leads to a substantial boost in performance, nearly doubling Chai-1's Pearson correlation and tripling it for AlphaFold3. As shown in Appendix A.3, the boost is consistent across complexes with single and multiple mutations. Figure 3 also shows per-structure correlations, where mutations are grouped by wild-type complex and metrics averaged within each group. By stratifying mutant groups by wild-type ipTM scores, we find that for high-quality complexes, pTMEnergy outperforms ipTM (further demonstrated in Appendix A.2). This indicates that confident structure predictions yield logit distributions that effectively capture binding-relevant energetic information.

4.2 ENERGY IDENTIFIES SUCCESSFUL RNA APTAMERS

Experimental setup. Next, we evaluate our EBMs on RNA aptamer virtual screening. We aim to identify aptamers that bind GFP from a large candidate pool. The dataset, curated by Huang et al. (2024), consists of GFPapt mutants (Shui et al., 2012). K_d values range from 0nM to 125nM, and aptamers with $K_d < 10$ are considered positives. We compare with all baselines reported in Huang et al. (2024) and structures are predicted using RosettaFold2NA (RF2NA) (Baek et al., 2024).

Table 1: Performance of energy-based scoring for RNA aptamer virtual screening. RosettaFold2NA results are averaged across five structure predictions per sequence.

		AUPRC	Precision@10	Precision@50
Baselines	Transformer	$.288_{.018}$	$.233_{.094}$	$.273_{.073}$
	SE(3) Transformer	$.288_{.015}$	$.200_{.141}$	$.267_{.061}$
	Equiformer	$.311_{.005}$	$.300_{.000}$	$.367_{.024}$
	EGNN	$.267_{.047}$	$.340_{.081}$	$.308_{.013}$
	GVP-GNN	$.317_{.071}$	$.300_{.081}$	$.380_{.082}$
	FA	$.290_{.014}$	$.300_{.141}$	$.333_{.033}$
	FAFormer	$.322_{.004}$	$.400_{.078}$	$.413_{.041}$
RosettaFold2NA	ipTM	$.253_{.001}$	$.150_{.035}$	$.290_{.007}$
	pAEnergy	$.343_{.006}$	$.400_{.071}$	$.380_{.000}$
	pTMEnergy	$\mathbf{.352_{.007}}$	$.400_{.071}$	$\mathbf{.420_{.000}}$

Benchmark Results. Table 1 shows that pTMEnergy achieves the highest AUPRC, outperforming the best structure-based baseline (FAFormer) by nearly 10%. pAEnergy also improved performance over RF2NA ipTM, highlighting its ability to better distinguish high-affinity binders.

4.3 Energy Identifies Miniprotein Binders

Experimental setup. Finally, we evaluate the accuracy of our EBMs in screening for protein-binding miniproteins. Our task is to distinguish miniprotein binders from nonbinders across multiple targets that were experimentally screened at large scale by Cao et al. (2022), focusing on the 10 targets selected by Bennett et al. (2023). Further dataset details are given in Appendix A.5. All metrics are computed using the Chai-1 model (no MSA), as AlphaFold3's MSA computation is prohibitively time-intensive. We benchmark against DSMBind (Jin et al., 2023) and FoldX (Schymkowitz et al., 2005) which is a physics-based energy function, using Chai-predicted structures for both.

Benchmark Results. As shown in Table 2, pTMEnergy achieves the strongest performance across all evaluation metrics. Further evidence that pTMEnergy, as a continuous energy-based metric, leads to better separation between binders and non-binders is given in Appendix A.6.

Table 2: Performance of different metrics in identifying miniprotein binders, averaged across targets.

Metric	AUPRC	Precision@10	Precision@50
FoldX	0.161	0.207	0.132
DSMBind	0.132	0.000	0.008
ipTM	0.177	0.300	0.244
pAEnergy	0.121	0.120	0.120
pTMEnergy	**0.181**	**0.340**	**0.252**

5 Conclusion

In this work, we demonstrate that protein structure predictors implicitly define energy-based models, enabling unsupervised binding energy prediction. By leveraging the confidence classification heads of structure predictors, we introduce pAEnergy and pTMEnergy which provide continuous measures of binding affinity. Our experiments show that these energy functions outperform traditional confidence metrics and unsupervised baselines across multiple tasks.

Our method poses some limitations. First, our framework requires running a protein folding model which can be computationally expensive. While Chai-1 in single-sequence mode offers a significantly faster alternative to MSA-based structure prediction, this computational cost remains a bottleneck for large-scale virtual screening applications. Future work could explore optimizations to accelerate structure prediction or approximate its outputs with a lighter-weight alternative.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

Minkyung Baek, Ryan McHugh, Ivan Anishchenko, Hanlun Jiang, David Baker, and Frank DiMaio. Accurate prediction of protein–nucleic acid complexes using rosettafoldna. *Nature methods*, 21(1): 117–121, 2024.

Nathaniel R Bennett, Brian Coventry, Inna Goreshnik, Buwei Huang, Aza Allen, Dionne Vafeados, Ying Po Peng, Justas Dauparas, Minkyung Baek, Lance Stewart, et al. Improving de novo protein binder design with deep learning. *Nature Communications*, 14(1):2625, 2023.

Longxing Cao, Brian Coventry, Inna Goreshnik, Buwei Huang, William Sheffler, Joon Sung Park, Kevin M Jude, Iva Marković, Rameshwar U Kadam, Koen HG Verschueren, et al. Design of protein-binding proteins from the target structure alone. *Nature*, 605(7910):551–560, 2022.

Feixiong Cheng, Junfei Zhao, Yang Wang, Weiqiang Lu, Zehui Liu, Yadi Zhou, William R Martin, Ruisheng Wang, Jin Huang, Tong Hao, et al. Comprehensive characterization of protein–protein interactions perturbed by disease mutations. *Nature genetics*, 53(3):342–353, 2021.

Chai Discovery, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhnikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, pp. 2024–10, 2024.

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pp. 2021–10, 2021.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

Brian L Hie, Varun R Shanker, Duo Xu, Theodora UJ Bruun, Payton A Weidenbacher, Shaogeng Tang, Wesley Wu, John E Pak, and Peter S Kim. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, 42(2):275–283, 2024.

Tinglin Huang, Zhenqiao Song, Rex Ying, and Wengong Jin. Protein-nucleic acid complex modeling with frame averaging transformer. *arXiv preprint arXiv:2406.09586*, 2024.

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, 2019.

Xiaoran Jiao, Weian Mao, Wengong Jin, Peiyuan Yang, Hao Chen, and Chunhua Shen. Boltzmann-aligned inverse folding model as a predictor of mutational effects on protein-protein interactions. *arXiv preprint arXiv:2410.09543*, 2024.

Wengong Jin, Xun Chen, Amrita Vetticaden, Siranush Sarzikova, Raktima Raychowdhury, Caroline Uhler, and Nir Hacohen. Dsmbind: Se (3) denoising score matching for unsupervised binding energy prediction and nanobody design. *bioRxiv*, pp. 2023–12, 2023.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Wei Lu, Jixian Zhang, Jihua Rao, Zhongyue Zhang, and Shuangjia Zheng. Alphafold3, a secret sauce for predicting mutational effects on protein-protein interactions. *bioRxiv*, pp. 2024–05, 2024.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021.

Bill R Miller III, T Dwight McGee Jr, Jason M Swails, Nadine Homeyer, Holger Gohlke, and Adrian E Roitberg. Mmpbsa. py: an efficient program for end-state free energy calculations. *Journal of chemical theory and computation*, 8(9):3314–3321, 2012.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan N Gomez, Debora Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pp. 16990–17017. PMLR, 2022.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International Conference on Machine Learning*, pp. 8844–8856. PMLR, 2021.

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(suppl_2):W382–W388, 2005.

Bo Shui, Abdullah Ozer, Warren Zipfel, Nevedita Sahu, Avtar Singh, John T Lis, Hua Shi, and Michael I Kotlikoff. Rna aptamers that functionally interact with green fluorescent protein and its derivatives. *Nucleic acids research*, 40(5):e39–e39, 2012.

Vinicius Zambaldi, David La, Alexander E Chu, Harshnira Patani, Amy E Danson, Tristan OC Kwan, Thomas Frerix, Rosalia G Schneider, David Saxton, Ashok Thillaisundaram, et al. De novo design of high-affinity protein binders with alphaproteo. *arXiv preprint arXiv:2409.08022*, 2024.

A APPENDIX

A.1 RELATED WORK

A variety of computational methods have been developed for binding energy prediction. We focus on unsupervised methods that do not rely on experimental labels which can be categorized into sequence-based, crystal structure-based, and structure prediction-based techniques.

Sequence-based. Recently, protein language models like ESM-1v (Meier et al., 2021) have enabled unsupervised prediction of protein mutation effects, enabling success for tasks like antibody affinity maturation (Hie et al., 2024). We compare the performance of these sequence-based models against our proposed energy-based scoring functions.

Crystal structure-based. DSMBind (Jin et al., 2023) is an energy-based model that estimates the likelihood of a particular crystal structure via SE(3) denoising score matching, performing well for protein-protein and antibody-antigen binding affinity prediction. BA-Cycle (Jiao et al., 2024) infers binding affinity from protein inverse folding model log-likelihoods, achieving state-of-the-art performance on PPI mutation effect prediction. However, for large virtual screening libraries, crystal structures will not be available.

Structure prediction-based. Advancements in biomolecular structure prediction have facilitated novel energy estimation techniques. Over the past year, AlphaFold3 (Abramson et al., 2024), followed by Chai-1 (Discovery et al., 2024), introduced a diffusion-based strategy that significantly enhances the accuracy of predicting multimeric complexes, including proteins, nucleic acids, and small molecules. AlphaFold3's ipTM metric has demonstrated signal for predicting PPI mutation effects on a small, filtered test set (Lu et al., 2024) and can identify miniprotein binders (Bennett et al., 2023). Unsupervised models like FAFormer (Huang et al., 2024), which outputs a contact map given predicted monomer structures, are also useful for virtual screening. We compare against these baselines in our study.

A.2 IMPACT OF STRUCTURE PREDICTION QUALITY ON SKEMPI PERFORMANCE

To assess the effect of structure prediction confidence on our ability to predict mutational impacts using extracted energy scores, we stratified the SKEMPI dataset by ipTM. At four different thresholds, we assess prediction performance across mutant groups whose wild-type complex has an ipTM score greater than the threshold.

Table 3: Correlation metrics at different ipTM thresholds.

ipTM Threshold	**0.0**	**0.5**	**0.8**	**0.9**
# of Examples	7082	4101	2501	1216
ipTM: Overall Pearson	0.102	0.171	0.153	0.208
pTMEnergy: Overall Pearson	0.265	0.362	0.400	0.430
ipTM: Per-Structure Pearson	0.252	0.338	0.367	0.357
pTMEnergy: Per-Structure Pearson	0.230	0.348	0.406	0.404

We find that as the ipTM threshold increases, the performance of our pTMEnergy improves. The overall Pearson correlation increases from 0.265 to 0.430 as the ipTM threshold rises from 0.0 to 0.9. This pattern also holds at the per-structure level. In contrast, the ipTM score itself has limited predictive utility for $\Delta\Delta G$ values, with much lower correlations even at high thresholds. Our energy-based method consistently outperforms raw ipTM in capturing the mutational effects, especially in high-confidence structure regimes.

A.3 SKEMPI PERFORMANCE: SINGLE VS MULTIPLE MUTATIONS

We evaluate model performance on the SKEMPI dataset by comparing Spearman correlations for complexes with single versus multiple mutations. The results are summarized in Figure 4. Overall, we observe that pTMEnergy significantly improves performance for complexes containing both one or many mutations, with a larger boost for multi-mutation examples.

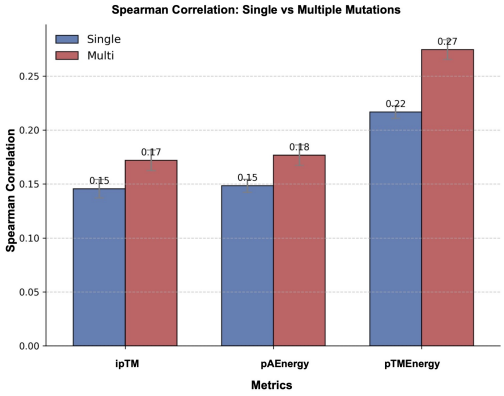

Figure 4: SKEMPI performance: single vs multiple mutations.

A.4 Energy Matrix Captures Mutational Effects

Figure 5 visualizes the difference in predicted pairwise residue interaction energy between the mutant and wild-type complex. The top row illustrates a deleterious mutation ($\Delta\Delta G > 0$), with the mutation site highlighted in green. The pAEnergy and pTMenergy functions assign a high interaction energy to the mutated position, reflecting the deleterious impact. Similarly, the bottom row depicts an advantageous mutation ($\Delta\Delta G < 0$) where the pAEnergy and pTMEnergy functions correctly assign lower predicted energy to the mutated position. We find that this trend holds consistently across the SKEMPI dataset, indicating that this continuous energy-based representation aligns with biophysical intuition, allowing for a more interpretable understanding of mutational effects.

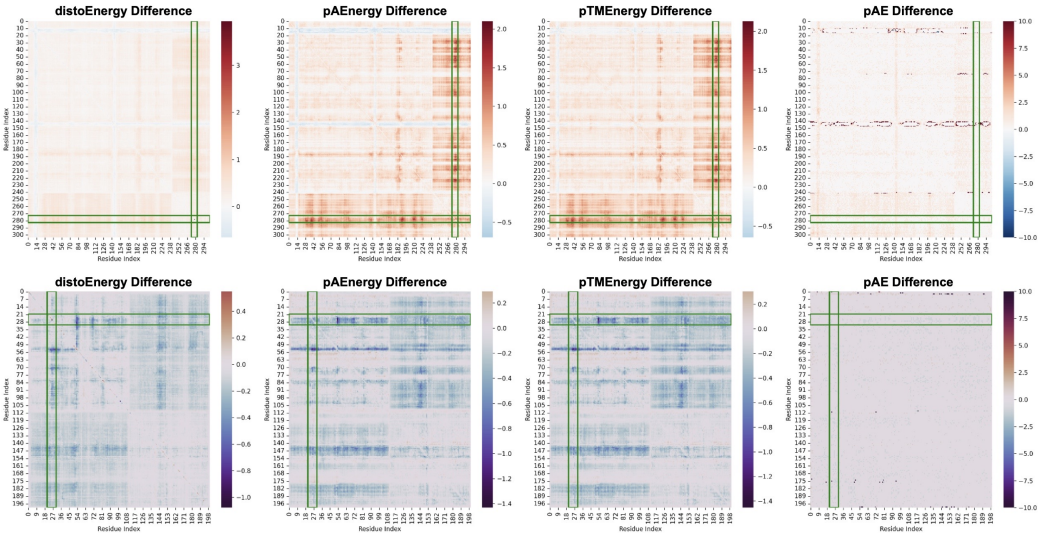

Figure 5: Difference between mutant and wild-type in AlphaFold3-predicted pairwise residue energy. Top row (PDB ID: 1ACB) illustrates a deleterious mutation while bottom row (PDB ID: 1B2U) shows an advantageous mutation. Mutation site is highlighted in green.

A.5 Miniprotein Dataset Details

We perform our analysis on 10 targets from Cao et al. (2022), as selected by Bennett et al. (2023). We subsample negative examples at a 10:1 negative-to-positive ratio.

Table 4: Miniprotein dataset details.

Target	Number of Positives
H3	50
TrkA	10
FGFR2	604
EGFR	15
PDGFR	284
InsulinR	259
Tie2	5
IL7Ra	22
VirB8	72
SARS-CoV2-RBD	18

A.6 MINIPROTEIN PERFORMANCE BY TARGET

For 8 out of 10 targets, pTMEnergy exhibits a higher Earth Mover's Distance between the binder and non-binder score distributions compared to ipTM. pTMEnergy, as a continuous energy-based metric, leads to better separation and captures finer variations in interaction strength.

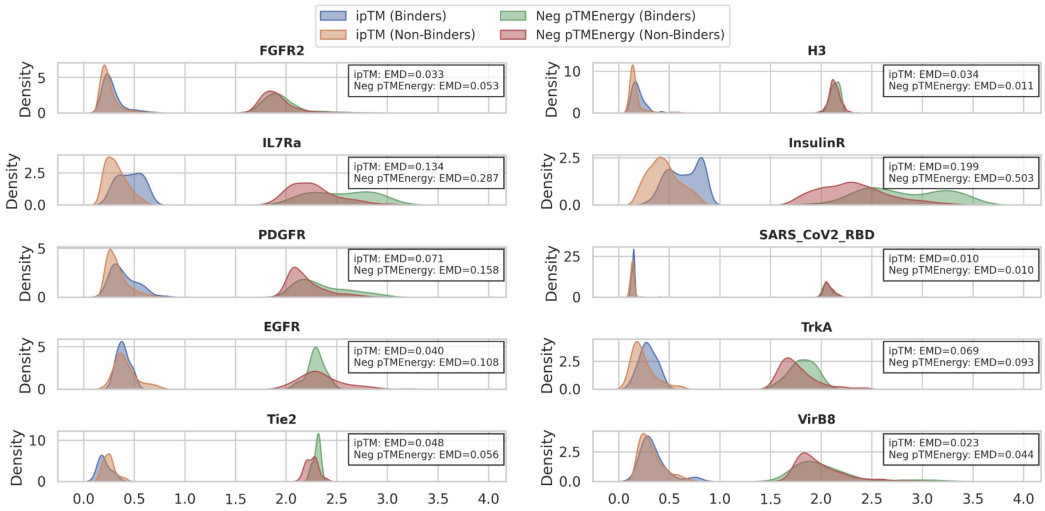

Figure 6: Distributions of ipTM and negative pTMEnergy for binders and non-binders across various binding targets. The Earth Mover's Distance (EMD) between binder and non-binder score distributions is consistently higher for pTMEnergy.