Improving scoring functions for protein-protein docking with LambdaLoss

Richard Zhu
Department of Statistics
Harvard University
rzhu@college.harvard.edu

Darren Xu, Lee-Shin Chu & Jeffrey J. Gray
Department of Chemical and Biomolecular Engineering
Johns Hopkins University
{dxu39,lchu11}@jh.edu, jgray@jhu.edu

Abstract

Modeling protein-protein interactions requires accurate scoring functions that can rank potential poses (conformations) of a protein-protein complex to differentiate near-native poses from incorrect ones. Here, we propose a general framework for improving protein-protein pose ranking models and other biomolecular interaction models using the LambdaLoss loss function from the Learning-to-Rank subfield. We test this framework by fine-tuning the energy prediction head of DFMDock with the LambdaLoss on an augmented dataset of 2.9M decoy poses derived from the DIPS dataset. On targets from the CAPRI score set benchmark, our fine-tuned ranking model LambdaDockScore is better at identifying correct poses in its top-1 and top-5 predictions compared to EuDockScore, a state-of-the-art method. LambdaDockScore also improves upon baseline DFMDock ranking performance for scoring antibody-antigen complexes and protein-protein complexes with very large or small binding interfaces.

1 Introduction

Protein-protein interactions are ubiquitous in biology, underlying many fundamental processes such as the recognition of viral antigens by human antibodies, the function of molecular machines like ATP synthase, and the conformational changes that drive signal transduction pathways. Recent advances in biomolecular modeling, such as AlphaFold3, have improved our capacity to model protein-protein interactions [1,12]. However, certain interactions remain challenging to model, such as predicting the interaction between antibody and antigen proteins [5,19].

Modeling protein-protein interactions is often decomposed into two subprocesses: 1. sampling plausible poses (conformations) of the protein-protein complex and then 2. ranking the generated poses to find the correct one, usually based on scoring functions that describe the biophysical quality of the poses. Thus, the development of accurate scoring functions is critical for success in modeling protein-protein interactions. Recently, deep-learning-based scoring functions have been shown to outperform classical methods and set a new bar in identifying near-native poses [13]. However, they can still struggle to correctly rank the poses of difficult protein-protein complexes [13].

In this study, we propose a novel framework for improving deep-learning-based scoring functions for biomolecular interactions. While we focus on protein-protein interactions here, this technique could be directly applied to other interactions as well. We employ a careful data augmentation process to generate a large dataset of protein-protein poses that spans a wide spectrum of biophysical quality, and then we fine-tune a base ranking model on our augmented dataset with the LambdaLoss. LambdaLoss is a machine learning loss function used in ranking query results in information retrieval systems like search engines [3,18].

We test our framework on the energy prediction head of DFMDock, a unified sampling and ranking model for protein-protein docking whose training objective encourages learning the underlying energy

Machine Learning for Structural Biology Workshop

landscape of protein-protein interactions [4]. We fine-tune this base model using the LambdaLoss and our augmented dataset to create the LambdaDockScore ranking model. On a set of 52 targets in the CAPRI score set [8], LambdaDockScore outperforms baseline DFMDock and a state-of-the-art ranking model EuDockScore in identifying the correct poses among the top-1 and top-5 ranked poses. We analyze LambdaDockScore's performance on subtypes of protein-protein complexes using another test set generated by running DFMDock inference on the Docking Benchmark 5.5 (DB5.5) [17]. In this test set, LambdaDockScore demonstrates the largest improvements relative to the baseline DFMDock ranking performance for two subtypes: antibody-antigen complexes and protein-protein complexes with very large or very small binding interfaces.

2 Related work

Please refer to the Supplementary Material for a summary of deep learning methods for modeling protein-protein interactions (Section A.1), scoring functions (Section A.2), and Learning-to-Rank methods like LambdaLoss (Section A.3).

3 Method

3.1 Applying LambdaLoss for biomolecular ranking

To adapt the LambdaLoss ranking framework to scoring biomolecular interactions, we modify the original LambdaLoss objective as follows. For each biomolecular complex (protein-protein complex, protein-nucleic acid complex, *etc.*) in the training set, we have a set of sampled poses $P = \{p_1, ..., p_N\}$. For pose $p_k \in P$, its ground truth quality is denoted as q_k, its predicted score (quality) is denoted as s_k, and its rank in the pose list sorted by q is denoted as $\mathrm{rank}(k)$. Then, we can define the biomolecular interaction LambdaLoss as the sum of pairwise loss terms between each pair of poses $p_i, p_j \in P$ for which $q_i > q_j$, as shown in Equation 1. This captures the rank- and score-weighted difference between the ground truth and predicted lists of pose rankings.

$$\mathcal{L} = \mathbb{E}\left[\sum_{q_i > q_j} \Delta_q \, \Delta_{\mathrm{rank}} \, \log\left(1 + e^{-(s_i - s_j)}\right)\right] \tag{1}$$

The weighting terms in Equation 1 are defined as

$$\Delta_q = \left|2^{q_i} - 2^{q_j}\right|, \qquad \Delta_{\mathrm{rank}} = \left|\frac{1}{\log(1 + \mathrm{rank}(i))} - \frac{1}{\log(1 + \mathrm{rank}(j))}\right| \tag{2}$$

Together, these equations define the biomolecular interaction LambdaLoss training objective.

3.2 Fine-tuning DFMDock

We fine-tune DFMDock's protein-protein scoring ability using the biomolecular interaction LambdaLoss in Equation 1. For this specific application, the ground truth quality metric q is the protein-protein complex pose's DockQ score, a measure of distance from the native complex structure [2]. The predicted score s is $-E$, where E is the energy output by DFMDock's energy prediction head. Thus, a lower energy corresponds to a higher predicted score and a prediction of higher pose quality. The DFMDock base model was chosen because its original training objective encourages it to learn the energy landscape of protein-protein docking. However, in the original training objective, the model only learns to optimize for the ground truth (the experimentally-determined native state), which is a single docking pose per protein-protein complex. By fine-tuning the model, we further refine the landscape by providing 270 non-native-state decoys per complex in the training dataset and applying the LambdaLoss to encourage the model to rank these decoys (and the ground truth pose) correctly by their DockQ scores. We refer to DFMDock before fine-tuning as the "baseline" model (which includes both a pose sampler and energy prediction head), while the fine-tuned energy prediction head is referred to as our "LambdaDockScore" model.

The 270 decoys per complex in the training set were generated according to the parameters in Supplementary Table S1. 220 complexes were generated by perturbing the ground truth protein-protein complex with varying levels of noise, while 50 complexes were generated as DFMDock predictions for the complex structure. The ground truth pose was also included for a total of 271 poses per complex. For each protein complex in the training set, we used this decoy sampling strategy to generate a set of poses spanning the $[0, 1]$ range of DockQ quality scores. This is shown in Supplementary Figure S1, where most poses with DockQ < 0.23 are generated from DFMDock sampling, while most poses with DockQ ≥ 0.23 are generated from perturbations of the ground truth.

This wide range in pose quality is essential for the success of the data-augmentation-based fine-tuning strategy for training LambdaDockScore, as it allows the model to rank a wide range of pose qualities encountered at inference time.

After generating the fine-tuning dataset, the entire DFMDock model is fine-tuned using its energy predictions and the LambdaLoss loss function shown in Equation 1. This loss function weighs the difference in predicted energies for two poses based on the difference in DockQ score (Δ_q) and the difference in rank when ranked by DockQ (Δ_{rank}). The loss function's sensitivity to DockQ difference helps LambdaDockScore differentiate between poses with very different scores, while its sensitivity to rank helps LambdaDockScore focus on the absolute best poses in the training set.

3.3 Dataset description and evaluation procedures

The final fine-tuning dataset contains 2,906,199 decoy poses for 10,724 protein-protein complexes from the DIPS-hetero training dataset. Validation was performed on a different subset of the DIPS-hetero dataset. Fine-tuning was conducted for 36 epochs with a learning rate of 1×10^{-4}. For each protein complex in each training epoch, ten decoys were sampled for ranking that spanned the entire range of DockQ scores. This was achieved by dividing the DockQ $[0, 1]$ range into ten buckets of width 0.1, assigning each pose a sampling probability of $\frac{1}{\text{\# poses in the bucket}}$, and sampling ten poses from the resulting distribution. On average, each bucket had one pose sampled, allowing the model to learn a wide range of DockQ scores.

For evaluation, LambdaDockScore was used either as a standalone scoring function for existing decoy poses (CAPRI score set) or paired with the baseline DFMDock sampling model for test datasets without existing decoy poses (DB5.5). In both cases, the top-1 and top-5 success rate metrics were calculated, which entails identifying the top-k decoy poses with the lowest predicted energy from the LambdaDockScore output and calculating their ground truth DockQ values. The top-k success rate is then the percent of protein-protein complexes in the test set where at least one of the top-k decoy poses ranked by LambdaDockScore surpasses the DockQ thresholds for acceptable- (DockQ > 0.23), medium- (DockQ > 0.49), or high-quality (DockQ > 0.80) poses, as defined by the Critical Assessment of PRedicted Interactions (CAPRI) initiative [2,7].

4 Results

4.1 LambdaDockScore outperforms a state-of-the-art protein-protein interface scoring function on the CAPRI score set

In Figure 1, we compare the scoring ability of LambdaDockScore (red) with the pose-scoring energy prediction head of the baseline DFMDock model (blue) and a state-of-the-art protein-protein ranking model EuDockScore (green). Since DFMDock and LambdaDockScore were trained on the DIPS-hetero subset of the DIPS dataset [11,16], while EuDockScore was trained on DB5.5 [17], the CAPRI score set was chosen as a challenging, independent evaluation set for all three models [8]. We use all complexes from the CAPRI score set (v2022) and performed additional filtering to eliminate sequence overlap with the DIPS-hetero training data for LambdaDockScore and DFMDock. By filtering for <30% sequence identity to any protein complexes in the DIPS-hetero training and validation sets, the CAPRI score set was reduced to 52 protein-protein complexes with an average of 1,037 decoy poses each for ranking. The three models were asked to rank all decoy poses for each protein-protein complex, and the DockQ scores of the top-1 and top-5 candidates were assessed.

As shown in Figure 1, LambdaDockScore outperforms the baseline DFMDock model at top-1 and top-5 ranking of acceptable and medium quality poses, though it underperforms or matches baseline

performance in ranking high quality poses. LambdaDockScore also outperforms state-of-the-art EuDockScore at top-1 and top-5 ranking of all three pose quality thresholds.

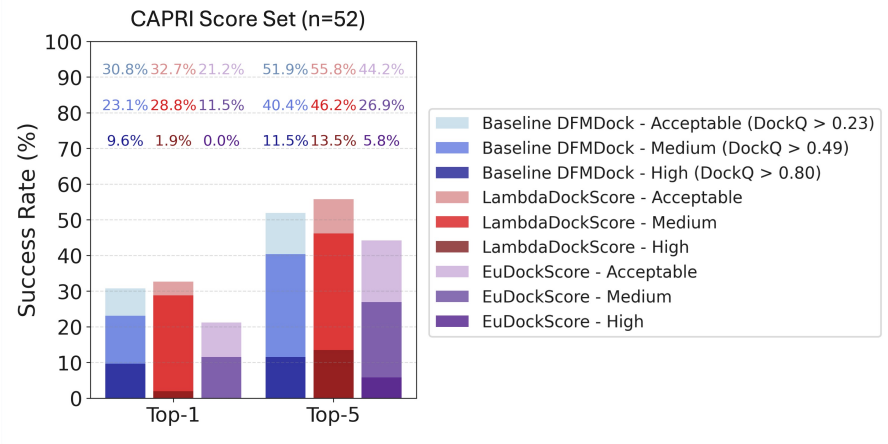

Figure 1: Benchmarking results on 52 protein-protein complexes from the CAPRI Score Set (filtered for <30% seq. identity to training data). Each complex averaged 1.4k decoys. The top-1 and top-5 ranked decoys from each model were analyzed.

4.2 LambdaDockScore improves docking performance on antibody-antigen complexes and protein complexes with extreme binding interface sizes

In Figure 2, we compare the pose ranking performance of LambdaDockScore and baseline DFMDock energy prediction head for docking poses sampled with the baseline DFMDock model on the DB5.5 dataset. A total of 120 decoy poses per protein-protein complex in the dataset were generated, and then either the baseline energy prediction head or LambdaDockScore was used to rank the poses. The top-1, top-5, and oracle success rates are calculated, where the oracle is the maximum possible success rate given a perfect ranking of the sampled poses.

LambdaDockScore outperforms baseline DFMDock energy prediction head's top-1 and top-5 performance for both acceptable and medium quality poses. Therefore, the performance of DFMDock can be improved by combining its sampling process with the fine-tuned LambdaDockScore for ranking.

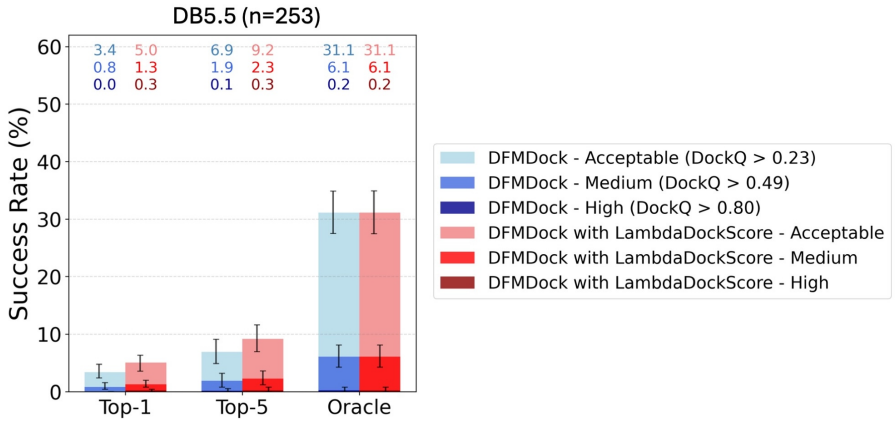

Figure 2: Benchmarking results on 253 protein-protein complexes from Docking Benchmark 5.5. 120 decoys were generated per protein-protein complex using baseline DFMDock. The top-1 and top-5 ranked decoys from each model were analyzed. DFMDock with LambdaDockScore as its ranking model outperforms baseline DFMDock on both top-1 and top-5 ranking accuracy for acceptable and medium quality poses. Bar heights show the mean performance from 10,000 bootstrap samples, and error bars show the 95% confidence interval.

Next, we analyze LambdaDockScore's performance on specific subtypes of protein-protein complexes within the DB5.5 benchmark set. Figure 3 shows that using LambdaDockScore improves baseline DFMDock's performance on antibody-antigen complexes in the DB5.5 benchmark dataset. Success rate improves from 1.2% to 1.9% for top-1 predictions and from 2.3% to 4.9% for top-5 predictions. Given the challenging nature of modeling antibody-antigen complexes, it is unsurprising that the absolute success rate percentage is low, as the oracle success rate (the maximum possible success rate) is only 17.6%. Since the success rate is a function of both the sampling performance (with baseline DFMDock) and ranking performance (with LambdaDockScore), using a better baseline sampling method than DFMDock may further improve the overall performance on these difficult protein-protein complexes.

Finally, Supplementary Figure S2 shows LambdaDockScore's performance on DB5.5 segmented by the size of the binding interface as measured by the change in accessible surface area upon binding (ΔASA). We show that the performance improvement for ranking acceptable and medium quality poses is the greatest for protein-protein complexes in the bottom 40% (Supplementary Figure S2a-b) and in the top 20% of interface sizes (Supplementary Figure S2e), while the performance drops in the middle 40% of protein-protein complex interface sizes (Supplementary Figure S2c-d). Thus, fine-tuning with LambdaLoss allows the pose ranking model to better rank protein-protein complexes with extreme-sized (very small or very large) binding interfaces, with the trade-off of decreased performance on average-sized binding interfaces.

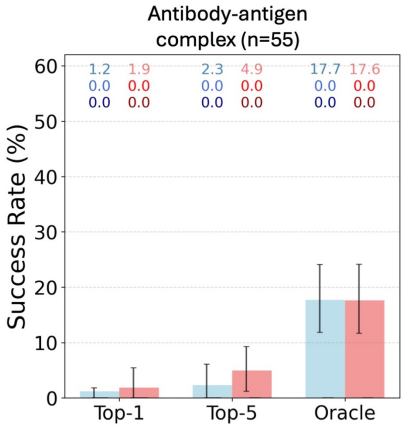

Figure 3: Benchmarking results on antibody-antigen complexes from Docking Benchmark 5.5. Results are for 120 decoys generated and ranked per protein-protein complex. DFMDock with LambdaDockScore as its ranking model (red) outperforms baseline DFMDock (blue).

5 Conclusion

In this study, we present LambdaDockScore, a novel method for improving protein-protein complex scoring functions using the LambdaLoss ranking loss. We show that by augmenting the original training data with carefully generated decoys and applying the LambdaLoss when fine-tuning the original ranking model on these decoys, we can not only improve the pose ranking accuracy beyond baseline, but also outperform EuDockScore, a state-of-the-art protein-protein complex scoring function. We also demonstrate that this data-augmentation-based LambdaLoss fine-tuning strategy specifically improves pose ranking performance on complexes with very large or very small binding interfaces, as well as antibody-antigen complexes. Improving performance on these challenging subcategories underscores the potential of LambdaDockScore. This LambdaLoss-based data augmentation and fine-tuning strategy could be used to improve numerous pose ranking problems and techniques in biomolecular modeling, such as the confidence modules of Boltz and AlphaFold. This would allow for improvements in not just protein-protein interactions, but also the grand challenge of modeling all biomolecular interactions.

6 References

[1] 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., … Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature, 630*(8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w

[2] Basu, S., & Wallner, B. (2016). DockQ: A Quality Measure for Protein-Protein Docking Models. *PLOS One, 11*(8), e0161879. https://doi.org/10.1371/journal.pone.0161879

[3] Burges, C. J. C., Ragno, R., & Le, Q. V. (2007). Learning to Rank with Nonsmooth Cost Functions. *Advances in Neural Information Processing Systems 19*, 193–200. https://doi.org/10.7551/mitpress/7503.003.0029

[4] Chu, L.-S., Sarma, S., & Gray, J. J. (2024). Unified Sampling and Ranking for Protein Docking with DFMDock. *bioRxiv*. https://doi.org/10.1101/2024.09.27.615401

[5] Hitawala, F. N., & Gray, J. J. (2025). What does AlphaFold3 learn about antibody and nanobody docking, and what remains unsolved? *mAbs, 17*(1), 2545601. https://doi.org/10.1080/19420862.2025.2545601

[6] Ketata, M. A., Laue, C., Mammadov, R., Stärk, H., Wu, M., Corso, G., Marquet, C., Barzilay, R., & Jaakkola, T. S. (2023, April 8). DiffDock-PP: Rigid Protein-Protein Docking with Diffusion Models. *The Eleventh International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.2304.03889

[7] Lensink, M. F., Méndez, R., & Wodak, S. J. (2007). Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins: Structure, Function, and Bioinformatics, 69*(4), 704–718. https://doi.org/10.1002/prot.21804

[8] Lensink, M. F., & Wodak, S. J. (2014). Score_set: A CAPRI benchmark for scoring protein complexes: A Benchmark for Scoring Protein Complexes. *Proteins: Structure, Function, and Bioinformatics, 82*(11), 3163–3169. https://doi.org/10.1002/prot.24678

[9] Liu, T.-Y. (2009). Learning to Rank for Information Retrieval. *Foundations and Trends® in Information Retrieval, 3*(3), 225–331. https://doi.org/10.1561/1500000016

[10] McFee, M., Kim, J., & Kim, P. M. (2024). EuDockScore: Euclidean graph neural networks for scoring protein–protein interfaces. *Bioinformatics, 40*(11). https://doi.org/10.1093/bioinformatics/btae636

[11] Morehead, A., Chen, C., Sedova, A., & Cheng, J. (2023). DIPS-Plus: The enhanced database of interacting protein structures for interface prediction. *Scientific Data, 10*(1), 509. https://doi.org/10.1038/s41597-023-02409-3

[12] Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., Kwabi-Addo, D., Beaini, D., Jaakkola, T., & Barzilay, R. (2025). Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. *bioRxiv*. https://doi.org/10.1101/2025.06.14.659707

[13] Shirali, A., Stebliankin, V., Karki, U., Shi, J., Chapagain, P., & Narasimhan, G. (2025). A comprehensive survey of scoring functions for protein docking models. *BMC Bioinformatics, 26*(1), 25. https://doi.org/10.1186/s12859-024-05991-4

[14] Stebliankin, V., Shirali, A., Baral, P., Shi, J., Chapagain, P., Mathee, K., & Narasimhan, G. (2023). Evaluating protein binding interfaces with transformer networks. *Nature Machine Intelligence, 5*(9), 1042–1053. https://doi.org/10.1038/s42256-023-00715-4

[15] Sverrisson, F., Feydy, J., Correia, B. E., & Bronstein, M. M. (2021). Fast end-to-end learning on protein surfaces. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15267–15276. https://doi.org/10.1109/CVPR46437.2021.01502

[16] Townshend, R. J. L., Bedi, R., Suriana, P. A., & Dror, R. O. (2019, December 26). End-to-End Learning on 3D Protein Structure for Interface Prediction. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. https://doi.org/10.48550/arXiv.1807.01297

[17] Vreven, T., Moal, I. H., Vangone, A., Pierce, B. G., Kastritis, P. L., Torchala, M., Chaleil, R., Jiménez-García, B., Bates, P. A., Fernandez-Recio, J., Bonvin, A. M. J. J., & Weng, Z. (2015). Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *Journal of Molecular Biology, 427*(19), 3031–3041. https://doi.org/10.1016/j.jmb.2015.07.016

[18] Wang, X., Li, C., Golbandi, N., Bendersky, M., & Najork, M. (2018). The LambdaLoss Framework for Ranking Metric Optimization. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 1313–1322. https://doi.org/10.1145/3269206.3271784

[19] Yin, R., & Pierce, B. G. (2024). Evaluation of AlphaFold antibody–antigen modeling with implications for improving predictive accuracy. Protein Science, 33(1), e4865. https://doi.org/10.1002/pro.4865

A Supplementary Material

A.1 Modeling protein-protein interactions with deep learning

Diffusion-based deep learning models have shown great promise in protein-protein docking, which aims to predict the structures of bound protein complexes from the structures of unbound monomers. For example, DiffDock-PP is a diffusion model that performs rigid-body docking: it performs only translations and rotations (with no protein backbone or side chain movements) to predict the bound complex. DiffDock-PP set the baseline accuracy for deep learning methods on the Database of Interacting Protein Structure (DIPS) dataset [6]. Building on this, DFMDock proposed the first diffusion-based docking model to unify sampling and ranking in one framework: it simultaneously learns translation and rotation forces to sample poses of the bound complex and an energy function to score these poses. DFMDock demonstrates better generalization than DiffDock-PP with superior performance on standard benchmarks (*i.e.* DB5.5) and a learned energy function that exhibits realistic binding funnels [4]. Most diffusion-based docking methods are paired with separate confidence models for ranking [6]; DFMDock's innovation lies in using its own energy prediction head for ranking, eliminating the need for an external scoring model [4].

In parallel, large "co-folding" models have emerged that jointly predict the complex structure of two interacting proteins from their sequences. AlphaFold3 uses a diffusion module for structure prediction and can model interactions between proteins, nucleic acids, small molecules, and ions [1]. Boltz-2 is another all-atom co-folding model that builds on the architecture of AlphaFold to incorporate additional functionalities like method, contact, and pocket conditioning [12]. These models outperform diffusion-based docking models, setting the state-of-the-art in deep learning for protein-protein complex prediction. However, these co-folding approaches rely heavily on evolutionary information extracted from multiple sequence alignments (MSAs) of related proteins. When homologous proteins are scarce or the evolutionary process is unique—as in many antibody–antigen complexes—their performance degrades [5,19]. Like diffusion models, co-folding models generate candidate poses and then use internal confidence or scoring metrics to rank them; AlphaFold3 and Boltz-2 produce confidence metrics like pLDDT and PAE [1,12].

Thus, virtually all current methods sample protein-protein complex poses and use a separate confidence model to predict the biophysical quality of poses for ranking. In the case of DFMDock, the learned energy from its energy prediction head is used for ranking instead of a separate model.

A.2 Scoring functions

Scoring functions take a proposed protein-protein complex and predict its biophysical quality (*e.g.* how close it is to the native complex). In docking pipelines, this is typically the second step after sampling poses: the scoring function should assign high scores to near-native poses and low scores to incorrect ones.

In recent years, deep learning approaches have outperformed classical scoring functions on multiple protein-protein docking benchmarks [13]. Current state-of-the-art deep learning scoring models include PIsToN (Protein Interfaces with Transformer Network), which crops protein-protein interface patches into images and processes them with a vision transformer [14]; dMaSIF, which represents the protein surface as a point cloud and applies geodesic convolutions to compute an interface binding score [15]; and EuDockScore, which uses an SE(3)-equivariant graph neural network to score protein–protein interfaces [10]. These methods learn to use geometric, chemical and/or energy features via neural networks, allowing them to set a new standard in scoring functions for protein-protein docking.

A.3 Learning-to-Rank methods

LambdaLoss originates from the Learning-to-Rank field, where models are trained to optimize ranking metrics (*e.g.* normalized discounted cumulative gain (NDCG)) rather than simple regression or classification losses [9]. The LambdaLoss loss function optimizes an upper bound on the NDCG metric [18]. Given a list of entities that must be ranked by their relevance to the query, the LambdaLoss analyzes pairs of entities, where the pairwise terms depend on both the ground-truth ranks and the difference in true relevance scores [18]. This formulation bridges rank-based and score-based ranking objectives and demonstrates state-of-the-art performance on Learning-to-Rank benchmarks [3,18].

A.4 Details on the generation process and DockQ distribution of the fine-tuning dataset

Decoy set	Translation std. dev. (Å)	Rotation std dev. (°)	# poses
Perturbed GT — small	0.1	1.0	20
Perturbed GT — medium	1.0	7.0	100
Perturbed GT — large	2.5	20.0	100
DFMDock samples	N/A	N/A	50
Total (per GT complex)			270

GT stands for ground truth. Standard deviations are for Normal distributions centered at 0, which were used to sample random translations and rotations for the perturbed GT poses.

Table S1: Decoy pose sampling details

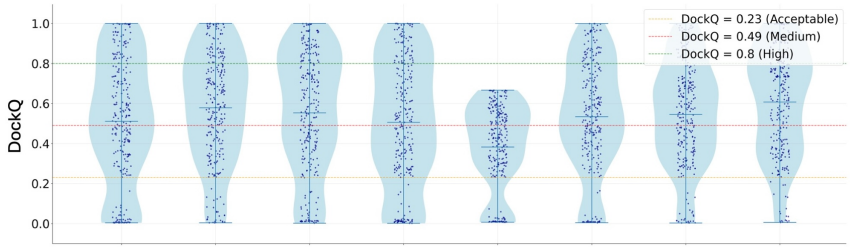

Figure S1: Distribution of poses for eight randomly sampled protein-protein complexes from the DIPS-hetero fine-tuning dataset. The majority of complexes in the fine-tuning dataset have a distribution of poses across the full DockQ [0,1] interval, though some complexes lack sufficient numbers of high DockQ poses.

A.5 LambdaDockScore performance on DB5.5, segmented by the size of the binding interface

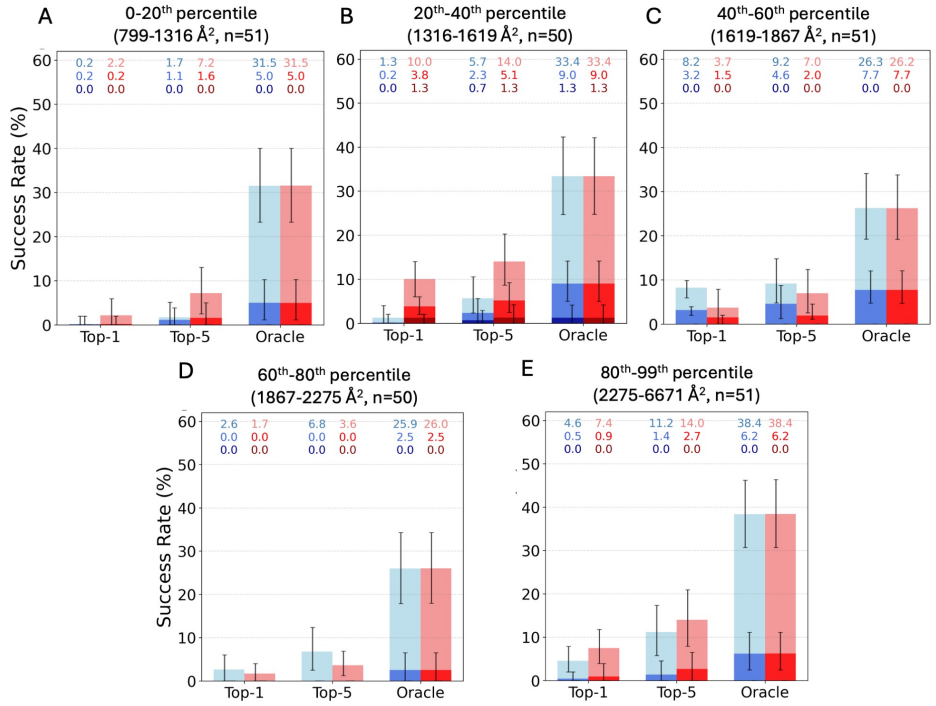

Figure S2: Benchmarking results on Docking Benchmark 5.5, segmented by the change in accessible surface area (ΔASA). This is a measure of how the model performs based on the size of the binding interface. Protein-protein complexes were ordered by ΔASA and split into five evenly sized bins based on their ΔASA percentile. Panels **(a)-(e)** show the performance of baseline DFMDock's energy prediction head (blue) compared to LambdaDockScore (red) for each bin, in order of increasing interface size.