Preprint, 2024, 1–10
doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

EuDockScore: euclidean graph neural networks for scoring protein-protein interfaces

Matthew McFee†,[1,2] Jisun Kim†[2] and Philip M. Kim[1,2,3]

[1]Department of Molecular Genetics, The University of Toronto, 1 King's College Cir., M5S 1A8, Ontario, Canada, [2]Donnelly Centre for Cellular and Biomolecular Research, The University of Toronto, 160 College St., M5S 3E1, Ontario, Canada and [3]Department of Computer Science, The University of Toronto, 40 St George St., M5S 2E4, Ontario, Canada

*Corresponding author: Philip M. Kim pi@kimlab.org
† indicates authors contributed equally

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Protein-protein interactions are essential for a variety of biological phenomena including mediating biochemical reactions, cell signaling, and the immune response. Proteins seek to form interfaces which reduce overall system energy. Although determination of single polypeptide chain protein structures has been revolutionized by deep learning techniques, complex prediction has still not been perfected. Additionally, experimentally determining structures is incredibly resource and time expensive, as well as difficult to perform. An alternative is the technique of computational docking, which takes the solved individual structures of proteins to produce candidate interfaces (decoys). Decoys are then scored using a mathematical function that predicts the energy of the system, know as scoring functions. Beyond docking, scoring functions are a critical component of assessing structures produced by many protein generative models. In this work we present improved scoring functions for protein-protein interactions which utilizes cutting-edge euclidean graph neural network architectures, in particular protein-protein docking scoring, as well as scoring antibody-antigen interfaces. These euclidean docking score models are known as EuDockScore, and EuDockScore-Ab with the latter being antibody-antigen specific. Additionally, we provide an antibody-antigen specific model specifically tuned to work with AlphaFold-Multimer outputs called EuDockScore-AFM showing evidence that the energy function learned by AlphaFold-like models can distilled. Finally, EuDockScore-AFSample is a model particular to scoring models from a the state-of-the-art complex predictor AFSample. These models can be used in conjunction with existing and new generative models to assess model energetics. The code for these models is available at `https://gitlab.com/mcfeemat/eudockscore`.

Key words: protein-protein docking, structural bioinformatics, deep learning, protein interaction, scoring function, energy function

1. Introduction

Protein-protein interactions are essential to a variety of biological processes including mediating chemical reactions, signaling, the immune response, and controlling the cell life cycle. Aberrations in protein-protein interactions are associated with variety of diseases including cancers (L. Garner and D. Janda, 2011). To understand these interactions, the structures of individual proteins as well as protein complexes need to be obtained and studied. Historically, these structures were solved for by using techniques such as x-ray crystallography, nuclear magnetic resonance, or cryogenic electronic microscopy. However, these techniques are expensive, time-consuming and challenging leading to the development of a variety of computational techniques to determine the interfaced structures of protein complexes (Zhang, 2008).

1.1. An introduction to protein-protein docking and scoring

Protein-protein docking encompasses the development of computational pipelines to propose potential interfaced complexes of two or more solved individual protein structures, most often from the Protein Data Bank (PDB) (Burley *et al.*, 2018). The docking

© The Author 2024. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

process is typically broken down into two stages. The docking procedure which proposes complexed structural models known as decoys, and the scoring stage (Tsuchiya *et al.*, 2022). The scoring stage involves a physical model taking in structural variables (e.g. coordinates and atom types) and assigning a score to the complex indicating energetic favourability. Typically, to align with the concept of energies in physics, lower scores indicate more energetically favorable systems. These scoring functions are used to rank proposed models for selection and further investigation (Guedes *et al.*, 2021).

1.2. An overview of scoring functions

In the past, scoring functions were typically hand-engineered to capitulate the known physics driving the folding and interactions of proteins. For illustrative purposes, a popular traditional scoring function is the Rosetta energy function (Alford *et al.*, 2017), which includes terms involving the electrostatics of charged atoms, bond lengths, Van der Waals forces etc. These models involved some model parameters being fit to available physical datasets (typically small). Although these models are useful, they often fail to capture the correct energetics of protein systems and fail to fully leverage the increasing amount of available biological data. Deep learning allows for the development of highly complex non-linear functions using available data for model parameter tuning. These models have revolutionized many fields such as natural language processing (NLP) and computer vision.

Many earlier efforts for protein-protein interface scoring with deep learning involved discretization of protein-protein interfaces into 3D image-like grids to be input into convolutional neural networks (Schneider *et al.*, 2021; Renaud *et al.*, 2021; Wang *et al.*, 2019). However, these networks are not rotationally and translationally invariant (rotating and translating input structures may change model outputs) and loss of information is an inherent aspect of discretization of the model inputs. Next, researchers moved towards using graph neural networks (GNNs) which utilize mathematical graph representations of proteins as model inputs (Wang *et al.*, 2021; Réau *et al.*, 2022). Here protein residues act as nodes in the graph and edges represent spatial relationships between said residues. The models were significantly more successful but still required some engineering including selecting or developing invariant node and edge features.

Continued efforts have been made to develop advanced graph neural network architectures to operate on graph representations at both the atom and residue levels for proteins. Modern architectures typically depend on equivariant mathematical operations derived from physics (Zitnick *et al.*, 2022; Liao and Smidt, 2023; Liao *et al.*, 2023) which can operate directly on input 3D coordinates without any necessary engineering. Typically these models will only need atom types and 3D coordinates to make predictions. Here equivariance means that any outputs of the model that are vectors will rotate appropriately with the inputs, furthermore, scalar outputs (such as a score) output but these models are typically invariant (do not change with rotated or translated inputs).

1.3. Scoring functions in the era of AlphaFold

The protein bioinformatics field was revolutionized by the publication of AlphaFold2 (Jumper *et al.*, 2021), a deep learning model which can take in single protein sequences, create a multiple sequence alignment, and use available template structures to generate exceptionally accurate folded protein models. The AlphaFold2 architecture was then extended to generate the structures of multi-chain protein complexes in the form of AlphaFold-Multimer (Evans *et al.*, 2022), providing an alternative avenue to computational docking techniques. Although AlphaFold-Multimer can produce accurate complexes models in a large number of cases, there are many inputs in which it still struggles to correctly determine the interface (Evans *et al.*, 2022). Finally, AFSample (Wallner, 2023), introduces new stochasticity to the AlphaFold-Multimer pipeline increasing the overall ability to recover a near-native complex model.

Other groups have attempted to use generative deep learning techniques to dock individual protein structures together with some success as well, particularly in small molecules (Stärk *et al.*, 2022; Corso *et al.*, 2023), with some extension to protein-protein docking (Ketata *et al.*, 2023). Although often much faster than traditional docking methods there are many cases in which these models do not outperform traditional methods or even have issues with incorrect model geometry (Buttenschoen *et al.*, 2024). Thus, docking and scoring remain relevant to the current field of structural bioinformatics, in particular, generative modeling.

1.4. The learned energy landscapes of AlphaFold-like deep learning architectures

Interestingly, works have shown that models such a AlphaFold2 are actually learning energy-like scoring functions (Roney and Ovchinnikov, 2022). Furthermore, groups have shown that combining scoring functions with AlphaFold-Multimer quality prediction output metrics may actually improve the filtering of protein-protein complex models. In particular, our previous model GDockScore (McFee and Kim, 2023) was used to screen AlphaFold-Multimer models for COVID-19 related interfaces (Tomezsko *et al.*, 2024). Thus, scoring and traditional docking remain relevant in the current realm of protein-protein interactions and structural bioinformatics.

1.5. The specific challenges of antibody-antigen binding physics

Antibodies remain an important area of study in the development of biologics for the treatment of a variety of diseases and deep learning techniques for the design and development of antibody binders to specific epitopes is currently a very active area of deep learning research (Kim *et al.*, 2023) with generative pipelines often using some form of scoring function. For example, AbDiffuser (Martinkus *et al.*, 2024), which generates antibody structures, still uses an additional scoring function to assess whether or not the generated structure will bind to a given target antigen. Furthermore, antibody interface scoring is particular difficult due to the immense amount of structural variation in heavy chain H3 loops with little human-detectable pattern (Marks and Deane, 2017). Thus, this challenging problem may better be tackled by creating a specialized scoring function rather than relying on generalized models.

2. Results

2.1. An Euclidean graph neural network architecture utilizing equivariant attention operations leveraging protein natural language model embeddings

EuDockScore, our latest scoring function, uses the equivariant attention-based architecture developed in the Equiformer and

EquiformerV2 models (Liao and Smidt, 2023; Liao *et al.*, 2023). These architectures have achieved state-of-the-art performance in many molecular graph learning tasks, showing increased expressivity in comparison to other graph-based architectures. In these architectures, nodes are initialized with scalar and vector features of orders l, where the dimensionality of the feature is given by $d = 2l + 1$, and there may be an arbitrary number of each of feature scalar or vectors. Different orders of features can equivariantly, or invariantly update each other via tensor product operations first introduced in (Thomas *et al.*, 2018). The tensor product operation is introduced in its full form in the Methods section of this manuscript but it should prove useful for readers to read the `e3nn` paper (Geiger and Smidt, 2022) on Euclidean networks. The Equiformer (Liao and Smidt, 2023; Liao *et al.*, 2023) introduces node level attention by generation attention values from the invariant scalar node features. Relative positional information is encoded by using the tensor product between node features and the spherical harmonics (Ledesma and Mewes, 2020) of relative positional vectors. Two major modifications were made to the EquiformerV2 architecture for this model. We moved from an all atom representation to a residue-level representation for computational efficiency. Secondly, we input complexes rather than single polypeptide chains, and finally we provide scalar NLP embeddings, discussed in the next section, to the model instead of just raw coordinates and atom/residue identities in the form of a one-hot encoding. Please consult the EquiformerV2 (Liao *et al.*, 2023) manuscript for schematic and further details about the model architecture.

2.2. An interacting protein-pair natural language model to supplement structure-based deep learning models

Large language models that implement the transformer architecture (Vaswani *et al.*, 2017) have revolutionized natural language processing and immediately were put to use in learning protein embeddings using the plethora of available sequencing data. In particular, the models introduced by Elnaggar et al. (Elnaggar *et al.*, 2022) and more recently Meta's ESM2 (Lin *et al.*, 2022) have been impactful to the field of protein bioinformatics. Single sequence protein folding models illustrate the power of these models sequence embeddings providing evidence that NLP models trained on sequence data are implicitly learning about protein folding physics, for example, ESMFold (Lin *et al.*, 2022). Furthermore, protein contact maps have been shown to be recoverable to some extend from NLP embeddings (Wang *et al.*, 2022). Thus, NLP models are learning important protein physical ideas which can be applied to the task of scoring input structures.

Similarly to DeepRank-GNN-esm (Xu and Bonvin, 2024), another recent scoring model, we sought to utilize NLP embeddings to supplement structure-based graph-based neural network models. We started by taking DistilProtBert (Geffen *et al.*, 2022), a distilled version of the original ProtBert model (Elnaggar *et al.*, 2022). Distilled models (Stanton *et al.*, 2021) are smaller versions of larger models that have been trained to replicate the outputs of larger NLP models at multiple points in the model architecture, distilling what is learned by the larger model into a smaller, less computationally taxing model. We chose this technique so we could fine-tune a protein language model on pairs of interacting protein sequences without requiring the immense computational resources typically needed to train LLMs. We took non-redundant interacting protein sequence pairs from BioGRID (Oughtred *et al.*, 2020), concatenated them together with a separating token and fine-tuned DistilProtBert to learn about protein-protein interaction physics. These NLP embeddings are projected down using equivariant linear layers implemented in the `e3nn` package (Geiger and Smidt, 2022) and used to initialize the scalar features of the EuDockScore models.

2.3. EuDockScore outperforms state-of-the-art models on the challenging CAPRI score_set for protein-protein docking

As in our previous work GDockScore (McFee and Kim, 2023), we first pre-trained the model using docking decoys generated from solved structures in the PDB. Training the model resulted in lower training and validation losses in comparison to GDockScore and training converged in roughly half the number of previously required training iterations. We also found that sub-sampling down the original pre-training sets resulted in no impact on model performance, so for this work we only used approximately 500000 decoys to train EuDockScore. We then used the refined decoys generated from BM5 (Vreven *et al.*, 2015) to train the original DeepRank model (Renaud *et al.*, 2021) to fine-tune the model to better understand backbone flexibility in docking and interface formation. The CAPRI score_set (Lensink and Wodak, 2014) still remains the gold-standard challenging test set for assessing protein-protein docking scoring functions and we again use it as our primary benchmarking set. We chose to benchmark EuDockScore against our previous model GDockScore, and the latest iteration of the DeepRank (Renaud *et al.*, 2021; Réau *et al.*, 2022) scoring model recently published in the literature known as DeepRank-GNN-esm (Xu and Bonvin, 2024) referred to solely as DeepRank for the remainder of this manuscript for brevity. DeepRank is an appropriate choice as EuDockScore it is also graph-based (though different in archictural design) and incorporates ESM2 protein language model embeddings (Lin *et al.*, 2022) as a major input feature. We used the standard command-line running procedure provided in the model documentation. To compare the models we used the metrics of area under the curve (AUC) of the receiver operating characteristic (ROC) known as the AUCROC as well as compared precision-recall curves (Fig. 1a,b). We found that in comparison to DeepRank both EuDockScore and GDockScore had significantly better AUCROCs ($p < 2.2e-16$) for both comparisons indicating EuDockScore more effectively can select out true positives without undesirable false positives. In comparison to each other, EuDockScore and GDockScore have similar AUCROCs. This indicates that there is a threshold where we get more true positives and less false positives in comparison to GDockScore. However, when applying statistical testing, the difference in AUCROCs is not statistically significant ($p = 0.2468$). However, we still use this as evidence of the improved discriminatory power of EuDockScore in comparison as although areas under curves may be similar the distribution of the "density" of this area may be different and important such as in the case of ROCs. We further sought to investigate how the scores of near-native and incorrect data points where distributed by EuDockScore using a violin plot (Fig. 1c). We can see from the violin plot that EuDockScore very confidently applies high scores to near-native decoys but is less confident (indicated by a larger spread of distribution of scores) in incorrect decoys. Logically, this is because binding is highly specific and inherently will have very precise features in comparison to the infinite number of incorrect interfaces.

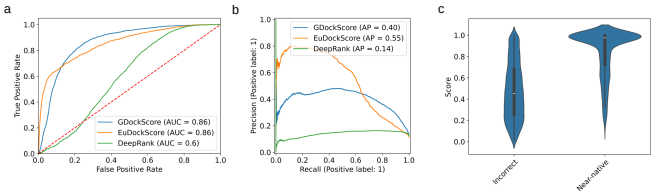

Fig. 1: (a) ROC curves comparing EuDockScore, GDockScore and DeepRank. (b) Precision-recall curves of EuDockScore, GDockScore, and Deeprank. (c) Violin plot comparing the distribution of scores produced by EuDockScore for the CAPRI score_set for near-native and incorrect decoys.

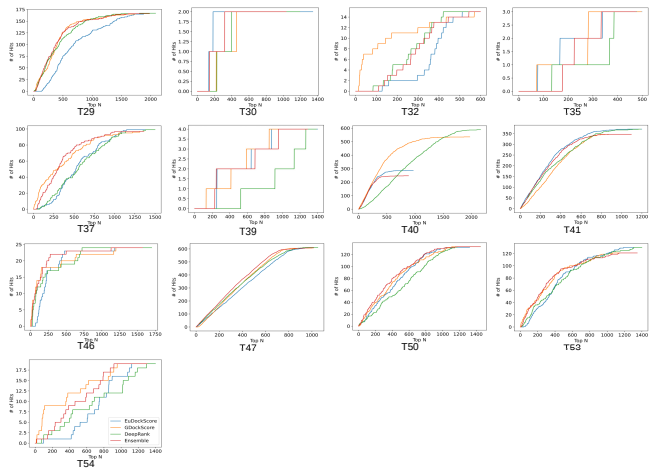

Fig. 2: Number of hits vs. top N selected curves for EuDockScore, GDockScore, DeepRank, and an ensemble average of all three models. See the Methods section for the definition of the model hit rate.

Furthermore, because in structural model assessment, biologists would only want to experimentally validate or study a handful of structure models they are confident are near-native, a scoring models ability to rank near-native decoys in the top N decoys is important. Typically small values of N are selected such as 5, or 10 but we choose to graphically display the number of hits in the top N selected structures (Figure 2), with the upper bound for N being the dataset size, as a more robust and comprehensive depiction of model ability to place near-native structures at high ranks. In Figure 2 below we can see that EuDockScore may do a better job on some structures (more near-native structures in a lower N) such as T30 and T40 than GDockScore but does worse on others. We also attempted to ensemble GDockScore, EuDockScore, and DeepRank to combine the classifying abilities of each model, since different models may work better on certain targets than others. Ensembling models for improved performance is a well studied machine learning phenomena (Dietterich, 2000). We found that ensembling all 3 models improved overall ranking ability (Figure 2).

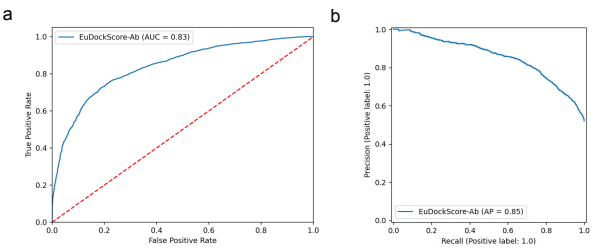

Fig. 3: (a) ROC curve for EuDockScore-Ab on hold-out, non-redundant antibody docks. (b) Precision-recall curve for EuDockScore-Ab on hold-out, non-redundant antibody docks.

2.4. EuDockScore-Ab: a novel antibody-antigen complex scoring model trained on antibody-antigen docks

Next, we sought to extend EuDockScore beyond standard protein-protein docking, into the realm of assessing antibody-antigen interfaces, a task still proving to be computationally very challenging. We fine-tuned the previous EuDockScore model to produce EuDockScore-Ab, an antibody specific model on a dataset of docked antigen-antibody structures (see Methods). To test the models ability to classify near-native from incorrect antibody-antigen complexes we used a non-redundant set of antibody structures generated by docking with RosettaDock (Gray *et al.*, 2003) and relaxation with the Rosetta protocol as well (Tyka *et al.*, 2011). Local and global docking was performed as in GDockScore (McFee and Kim, 2023) to ensure we captured incorrect and near-native decoys in the set. Using a hold out, non-redundant test set we see that EuDockScore-Ab can differentiate near-native from incorrect docks in unseen data (Fig. 3a,b). It is important to note our model is specific to scoring heavy chain and antigen complexes, as we found no benefit of training and using an additional light chain specific model, and scoring all three chains at once was often too computationally expensive.

2.5. EuDockScore-AFM: distilling the scoring abilities of AlphaFold-Multimer type architectures with ground-truth correction

We then sought to determine if the scoring abilities of EuDockScore-Ab extended to multimeric structures of antibodies and their antigens produced by other methods that classical docking and energy-based relaxation. We took AlphaFold-Multimer outputs generated by the ColabFold implementation (Mirdita *et al.*, 2022) and tested the performance of EuDockScore-Ab to differentiate decoys with acceptable or better DockQ scores vs. the output scoring method of AlphaFold-Multimer (0.8*ipTM + 0.2*pTM, dubbed the ranking confidence). We found on this set that although EuDockScore-Ab is worse in terms of AUCROC and AP of the precision-recall curve (Fig. 4a,b), EuDockScore-Ab does have predictive power on AlphaFold outputs and is not necessarily docking specific. For the AUCROCs EuDockScore-Ab is significantly different from AlphaFold-Multimer and the ensemble with ($p = 0.01817$, $p = 1.196e{-}07$) and the ensemble and AlphaFold-Multimer metrics due not have significantly different values ($p = 0.2694$).

Due to the implications that AlphaFold2 is learning an energy function (Roney and Ovchinnikov, 2022), the usefulness of distilling AlphaFold metrics (Melnyk *et al.*, 2024), and the

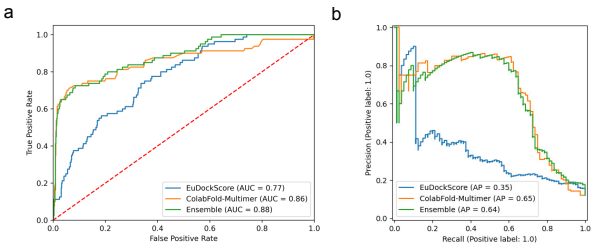

Fig. 4: (a) ROC curves comparing EuDockScore-Ab, ColabFold-Multimer (0.8*ipTM + 0.2*pTM), and an ensemble of both scoring metrics (b) Precision-recall curves comparing EuDockScore-Ab, ColabFold-Multimer (0.8*ipTM + 0.2*pTM), and an ensemble of both scoring metrics.

success of using GDockScore for filtering AlphaFold-Multimer outputs (Tomezsko *et al.*, 2024). we decided to explore developing a supplementary scoring model specifically trained for AlphaFold-Multimer antibody outputs. Instead of engineering AlphaFold2 to output single chain like scores by utilizing decoys as templates such as (Roney and Ovchinnikov, 2022) we propose that learning a new model directly on AlphaFold-Multimer outputs.

Instead of simply training to distill AlphaFold-Multimer, the new training target is the sum of the AFM score of the model and the DockQ against the native (true) structure. This combined metric allows our model to "correct" regions of the AlphaFold-Multimer energy function using ground truth structural data. We found that this training target of AlphaFold-Multimer model confidence + DockQ does in fact improve model performance on unseen test data. When taking this new model, hereby known as EuDockScore-AFM, and testing it on a hold-out test set of non-redundant antibody AlphaFold-Multimer structures, we find that our model performs very similarly to AlphaFold-Multimer quality metric (Fig. 5a,b) in both of terms of AUCROC and AP of the precision-recall curve. It is important to note that we only use the heavy chain and antigen as inputs to our model as before. The overall ranking ability of our model, as determined by hit rate, is similar to the AlphaFold-Multimer (Fig. 5c). Ensembling our model with the original AlphaFold-Multimer results in a small improvement in performance on our target metrics (Fig. 5a,b,c). The DeLong test indicates that the different between the AUCROCs of EuDockScore-AFM and the AlphaFold-Multimer metrics is not significant ($p = 0.5659$), and the difference between itself and the ensemble is also not significant ($p = 0.03547$) after Bonferonni correction. The ensemble and the AlphaFold-Multimer metric are not significantly different ($p = 0.6614$). Thus, we show evidence that we have to some extent distilled the inherent energy function learned by AlphaFold-Multimer in a model consisting of hundreds of thousands of parameters very tens of millions. Furthermore, we have multimeric scoring similar to AlphaFold-Multimer that allows for user input structures, allowing for greater applicability of scoring in comparison to AlphaFold-Multimer and any methods that modify AlphaFold to perform scoring (Roney and Ovchinnikov, 2022).

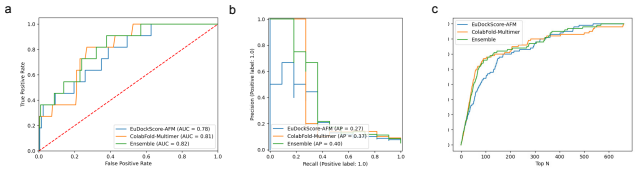

Fig. 5: (a) ROC curves comparing EuDockScore-AFM, ColabFold-Multimer (0.8*ipTM + 0.2*pTM), and an ensemble of both scoring metrics (b) Precision-recall curves comparing EuDockScore-AFM, ColabFold-Multimer, and an ensemble of both scoring metrics. (c) Number of near-native structures in top N selected structures curves for EuDockScore-AFM in comparison to ColabFold-Multimer and an ensembled model.

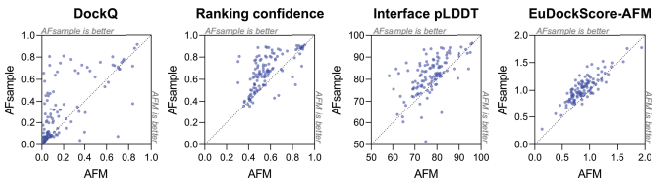

Fig. 6: Comparison of AFM and AFsample for antibody-antigen complex prediction. The best prediction was selected by each parameter indicated on the graph.

2.6. EuDockScore-AFSample proves to outperform AlphaFold-Multimer confidence metrics on selecting near-native complexes

Next, we confirmed whether or not EuDockScore-AFM can be used to score antibody-antigen complex structures predicted by AFSample, a new methodology building upon AlphFold-Multimer for complex predictions. With the ability to generate larger quantities of predicted structures through large-scale sampling and stochastic perturbation techniques, AFsample significantly outperformed the AlphaFold-Multimer baseline on the CASP15 benchmark, demonstrating improved accuracy in predicting complex protein structure (Wallner, 2023). Thus, we assumed that AFsample would show better performance for Antibody-antigen complex prediction than AlphaFold-Multimer. In the same way with AlphaFold-Multimer, we predicted complex structures with AFsample for all selected input antibody-antigen complexes taken from the PDB and analyzed them with EuDockScore-AFM and several existing parameters indicating the quality of complex structure (DockQ, ranking confidence, and interface pLDDT). AFsample showed better performance for antibody-antigen complex prediction than Alphafold-Multimer, in terms of every parameter we analyzed (Figure 7). When calculating the fractional hit rate and average DockQ, we suggested that EuDockScore-AFM is a useful indicator for selecting the best quality of predicted structure among 1000 decoys of AFsample (Table 1).

Next, we hypothesized that a custom scoring model may be useful in the case of AFSample (Wallner, 2023) which modifies the inference regime of AlphaFold-Multimer generating varied samples not typically generated by the original model. Thus, it may be beneficial to create a specialized scoring model to better understand the energetics of these new kinds of proposed models. Again we trained our model on the labels of AFSample

Parameter	DockQ	Ranking confidence	Interface pLDDT	EuDockScore-AFM
Success rate	47.9	22.7	21	23.5
Average DockQ	0.302	0.183	0.165	0.174

Table 1. Fractional hit rates and average DockQ for antibody-antigen complex predictions by AFsample. The best prediction was selected by each parameter. The fractional hit rate indicates the fraction of predicted structures that meet the acceptable criterion (DockQ $\geq$ 0.23) out of all unique model inputs.

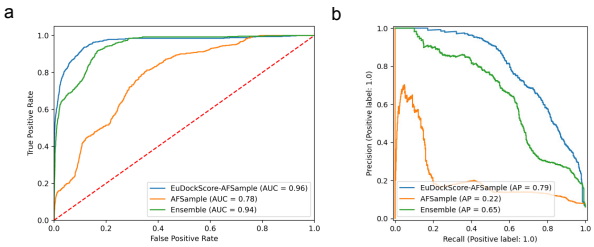

Fig. 7: (a) ROC for EuDockScore-AFSample in selecting near-native structures compared to AFSample confidence metrics, and an ensembled metric. (b) Precision-recall curve for EuDockScore-AFSample in selecting near-native structures compared to AFSample confidence metrics, and an ensembled metric.

Top N	10	50	100
EuDockScore-AFSample	0.321	0.375	0.393
AFSample	0.304	0.366	0.384
Ensemble	0.321	0.375	0.411

Table 2. Fractional hit rates for each model. Here fractional hit rate is the number of unique PDBs that had a near-native structure in the top N selected structures divided by the total number of unique PDB IDs.

score (same confidence metric as AlphaFold-Multimer) + DockQ values on a training set of AFSample generated structures, and tested on a non-redundant hold-out set. We found that our model outperforms AFSample confidence scores in terms of AUCROC and AP of the PR curve (7a,b). Here the AUCROC of EuDockScore-AFSample and AFSample is significantly different ($p < 2.2e-16$), EuDockScore-AFSample to Ensemble is also significant ($p = 1.236e-12$), and finally, AFSample and the ensemble are also significantly different ($p < 2.2e-16$). Therefore, we have learned a AFSample specific model that corrects for the deficiencies in AlphaFold-Multimer's confidence ranking on AFSample generations.

Similarly, to Figure 2, we sought to see how well our model ranks near-native structures in the top N selected scorers for each unique PDB ID in the hold-out test set. Since there are too many PDBs to show individual hit rate curves we display success rates in Table 2. Please consult the Methods section for a definition of success rates. For 3 top N values, we find that our model has a higher success rate than pure AFSample ranking confidences. However, in agreement with the docking results in Figure 2, we find that ensembling of the two scores with an average improves the overall success rate.

3. Discussion

Here we have presented several EuDockScore scoring functions that can be used to accurately score protein-protein docks as well as antibody-antigen complexes. We iterate upon our previous work (McFee and Kim, 2023) by using modern, increasingly expressive Euclidean graph networks that can operate on raw coordinates (Liao and Smidt, 2023; Liao *et al.*, 2023) and supplemented the inputs with NLP embeddings from a BERT-style model (Geffen *et al.*, 2022).

We continue to achieve state-of-the-art performance on the CAPRI score_set (Lensink and Wodak, 2014) and find that ensembling with other models such as DeepRank-GNN-esm (Xu and Bonvin, 2024) improves the overall discriminative power of available deep learning scoring models for docking.

To further test the abilities of our scoring framework to another challenging task, we trained antibody-antigen complex specific models known as EuDockScore-Ab and EuDockScore-AFM. The former was trained on relaxed docks of non-redundant antibody-antigen complexes, and the latter was trained on ColabFold-Multimer (Mirdita *et al.*, 2022) predictions of antibody-antigen complexes, with the AlphaFold-Multimer quality metric as a target. We find that EuDockScore-Ab can discriminate non-native vs native antibody docks, as well as ColabFold-Multimer predictions without training specifically on ColabFold-Multimer outputs. This suggests our model can to some extent recapitulate the physics learned by AlphaFold2 style models. We found that we could distill the energy function learned by AlphaFold2, circumventing the need to run the entire AlphaFold2 architecture, the inability to input user-defined structures, as well as working on multimeric structures in comparison to AF2Rank (Roney and Ovchinnikov, 2022).

Finally, we developed an AFSample-specific scoring model that outperforms the confidence metrics generated for models by AFSample. This illustrates the power of additional scoring models to supplement existing confidence metrics. The strong performance of our model is likely due to AFSamples unique samples having properties significantly different from the models generated by AlphaFold-Multimer and thus are not well classified by the AlphaFold-Multimer confidence metric models. Our model uses a summation of AlphaFold-Multimer confidence metrics and DockQs to help correct for this inadequacy.

Further work could be done to extend our presented results. For example, to generalize the distilled EuDockScore-AFM scoring function to general protein-protein interactions, AlphaFold-Multimer predictions for all multimers in the PDB could be generated and used to distill a generalized protein-protein model. This is also applicable to AFSample predictions.

4. Methods

4.1. Generation of an antibody-antigen structure dataset using docking

Due to our previous success with GDockScore and the pre-training regime we established (McFee and Kim, 2023). We decided to replicate this training scheme for an antibody-antigen complex dataset. We used the AbDb (Ferdous and Martin, 2018), an antibody database extracted from the PDB with comprehensive processing and redundancy analysis amounting to several thousand non-redundant antibodies. Although we simply fine-tuned the previous pre-trained EuDockScore on

the GDockScore and DeepRank decoys, we sought to increase antibody-specific training data complexity with re-docking and structural relaxation. To generate our dataset from AbDb, we performed local docking with the RosettaDock protocol (Gray *et al.*, 2003). In this local docking process small perturbations are applied to ligand (antibody) and the docking protocol is run. We used the CAPRI standard assessment for docking quality as specified in Mendez et al. (Méndez *et al.*, 2003). As in GDockScore (McFee and Kim, 2023) we assign non-native decoys as label 0 and acceptable or better decoys as label 1, or near-native. We generated roughly 50000 docking decoys across the AbDb with a 70%, 15%, 15% split to training, validation, and test sets. Redundancy information is provided by AbDb (Ferdous and Martin, 2018) and we ensured that no redundant structures were shared amongst the splits to ensure un-biased performance reporting. To further improve the quality of the interfaces and introduce back-bone flexibility into our docking models we ran the RosettaRelax protocol (Tyka *et al.*, 2011) on the docking decoys to relax the backbones of the structures and re-pack side chains as a more computationally efficient alternative to molecular dynamics simulations. Thus, we present a new dataset for training antigen-antibody specific scoring functions to the literature in line with our previously successful datasets.

For this dataset, to ensure no redundancy between train, validation, and test sets we used the redundancy information provided by the AbDb (Ferdous and Martin, 2018) to split the AbDb into train/validation/test splits.

4.2. Generation of an antibody-antigen complex dataset using AlphaFold-Multimer and AFSample

To generate AlphaFold-Multimer predictions of Ab complexes, we used the protocol above, but instead generated models using the local colabfold implementation of ColabFold (Mirdita *et al.*, 2022) with templates enabled and all other options set to the default settings. Again, we split the data to avoid redundancies by clustering antigens by 30% sequence similarity with MMSeqs2 (Steinegger and Söding, 2017) and making sure there was no cluster overlaps between the train, validation, and test set. Furthermore, the training and validation sets only contain PDBs released up to 2020 and the hold-out test set only contains PDBs released after 2020 in a similar fashion to the AlphaFold2 splitting (Jumper *et al.*, 2021) to ensure a fair comparison.

The AFSample data was generated as above with the default AFSample settings, and similarly split into non-redundant train/validation/test splits.

4.3. A mathematical description of the tensor product

Here we introduce the tensor product.

$$(u \otimes v)_m^{(l)} = \sum_{m_1=-l_1}^{l_1} \sum_{m_2=-l_2}^{l_2} C_{(l_1,m_1)(l_2,m_2)}^{(l,m)} u_{m_1}^{l_1} v_{m_2}^{l_2}$$

In this representation, u and v are irreducible representations of orders l_1 and l_2, where the dimension of u and v are given by the formula $d = 2l+1$. For example, scalar features would be $l = 0$ and $d = 2*0+1$. The C terms are Clebsch-Gordon coefficients (Griffiths and Schroeter, 2019). The m variables indicate the different indices of each quantity. This operation is provably equivariant or invariant. Clebsch-Gordon coefficients are non-zero for values between $|l_2 - l_1|$ and $(l_2 + l_1)$. Here v and u can be any valid node feature vectors or scalars or relative positional vectors encoded with spherical harmonics. For a comprehensive treatment of this operation consult the following works (Geiger and Smidt, 2022; Thomas *et al.*, 2018; Liao and Smidt, 2023; Liao *et al.*, 2023).

4.4. Details of the input representation

The inputs to EuDockScore are all the raw 3D coordinates of each amino acid residue, as well as one hot encoding of amino acid identity, a binary indicator if the amino acid is within 16 angstroms of an amino acid on the other protein, as well as associated amino acid embeddings from our fine-tuned DistilProtBert model. The scalar and vector input features are projected down using equivariant operations and used to initialize irreducible node features used in the Equiformer architectures (Liao and Smidt, 2023; Liao *et al.*, 2023).

4.5. Dataset redundancy analysis for BioGRID data

For the BioGRID dataset used to train the NLP model presented in this paper, we used the MCODE algorithm (Bader and Hogue, 2003) to cluster BioGRID and then split the data such that there was no cluster overlap in the interactions in the training and validation set.

4.6. Model implementation, hyperparameter tuning and training details

Hyperparameter tuning was performed by adjusting the number of spherical channels ie. number of types of each order l feature for each node in the EquiformerV2 (Liao *et al.*, 2023) architecture as well as adjusting the maximum channel order of the type-l features of the model. We would that lmax = 2 and 16 spherical channels. Three EquiformerV2 layers are used in the model. The optimal hyperparameter configuration is set to the default configuration in the model implementation. Furthermore, we trained the model using the Adam optimizer (Kingma and Ba, 2015) for approximately 70000 iterations with a batch size of 5. The model was implemented in PyTorch (Ansel *et al.*, 2024) and the ADAM optimizer was used to train the model (Kingma and Ba, 2015).

The target labels for the docking training madels were the same as GDockScore (McFee and Kim, 2023) and the loss was the binary cross-entropy function. The target labels for all AFM and AFSample specific models were the generating models confidence metric (0.8*iPTM + 0.2*PTM) + the DockQ value of the model against the native structure. The loss function to train these models was mean-squared error (MSE).

4.7. Definition of average precision

The average precision reported is from the implementation provided by the scikit-learn package (Pedregosa *et al.*, 2011). The mathematical definition is

$$\mathrm{AP} = \sum_n (R_n - R_{n-1}) P_n \tag{1}$$

Here AP represents average precision, the summation is over the n thresholds, R_n is the recall at the given threshold, R_{n-1} at the previous threshold, and the weighting term P_n is current precision.

4.8. Data analysis and statistical Testing

The ROCs precision-recall curves and other analysis presented in this paper were conducted using scikit-learn (Pedregosa *et al.*, 2011). The AUC of the ROC curves was compared using the DeLong test (DeLong *et al.*, 1988) provided by the pROC R (R Core Team, 2021) package (Robin *et al.*, 2011). The Bonferonni correction was used for multiple comparisons.

5. Competing interests

PMK is a co-founder and consultant to several biotechnology companies, including Fable Therapeutics, Resolute Bio, and Navega Therapeutics. PMK also serves on the scientific advisory board of ProteinQure. MM reports no competing interests. JK reports no competing interests.

6. Author contributions statement

The model architecture was conceptualized by MM and PMK. Model implementation was conducted by MM. Data generation and data analysis were performed by MM and JK. Writing was performed by MM, JK, and PMK. Supervision and funding acquisition was conducted by PMK.

7. Acknowledgments

We would like to thank the Digital Research Alliance of Canada for providing computing resources. We also acknowledge CIHR Project Grant PJT-159750 for funding this work.

8. Data and code availability

The data set introduced in this paper is available at `http://gdockscore.ccbr.proteinsolver.org` including the language model weights/data splits, as well as the new antibody-antigen docking decoy dataset. The CAPRI score set files used in this paper are available at `https://data.sbgrid.org/dataset/684/`. The DeepRank data set is available at `https://data.sbgrid.org/dataset/843/`. The original AbDb data is available at `http://www.abybank.org/abdb/`. The code is available at `https://gitlab.com/mcfeemat/eudockscore`.

References

Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T., and Gray, J. J. (2017). The rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation*, **13**(6), 3031–3048.

Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., Hirsh, B., Huang, S., Kalambarkar, K., Kirsch, L., Lazos, M., Lezcano, M., Liang, Y., Liang, J., Lu, Y., Luk, C. K., Maher, B., Pan, Y., Puhrsch, C., Reso, M., Saroufim, M., Siraichi, M. Y., Suk, H., Zhang, S., Suo, M., Tillet, P., Zhao, X., Wang, E., Zhou, K., Zou, R., Wang, X., Mathews, A., Wen, W., Chanan, G., Wu, P., and Chintala, S. (2024). Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS '24, page 929–947, New York, NY, USA. Association for Computing Machinery.

Bader, G. D. and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**(1), 2.

Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K., Duarte, J. M., Dutta, S., Feng, Z., Ghosh, S., Goodsell, D. S., Green, R. K., Guranović, V., Guzenko, D., Hudson, B. P., Kalro, T., Liang, Y., and Lowe, R. (2018). Rcsb protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Research*, **47**(D1), D464–D474.

Buttenschoen, M., Morris, G. M., and Deane, C. M. (2024). Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*.

Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. S. (2023). Diffdock: Diffusion steps, twists, and turns for molecular docking. In *The Eleventh International Conference on Learning Representations*.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, **44**(3), 837.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. (2022). Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**(10), 7112–7127.

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J., and Hassabis, D. (2022). Protein complex prediction with alphafold-multimer. *bioRxiv*.

Ferdous, S. and Martin, A. C. R. (2018). Abdb: antibody structure database—a database of pdb-derived antibody structures. *Database*, **2018**.

Geffen, Y., Ofran, Y., and Unger, R. (2022). Distilprotbert: a distilled protein language model used to distinguish between real proteins and their randomly shuffled counterparts. *Bioinformatics*, **38**(Supplement$_2$), *ii*95–*ii*98.

Geiger, M. and Smidt, T. (2022). e3nn: Euclidean neural networks.

Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., and Baker, D. (2003). Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology*, **331**(1), 281–299.

Griffiths, D. J. and Schroeter, D. F. (2019). *Introduction to Quantum Mechanics*. Cambridge University Press.

Guedes, I. A., Barreto, A. M. S., Marinho, D., Krempser, E., Kuenemann, M. A., Sperandio, O., Dardenne, L. E., and Miteva, M. A. (2021). New machine learning and physics-based scoring functions for drug discovery. *Scientific Reports*, **11**(1).

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., and Back, T. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, **596**(7873), 583–589.

Ketata, M. A., Laue, C., Mammadov, R., Stärk, H., Wu, M., Corso, G., Marquet, C., Barzilay, R., and Jaakkola, T. S. (2023). Diffdock-pp: Rigid protein-protein docking with diffusion models.

Kim, J., McFee, M., Fang, Q., Abdin, O., and Kim, P. M. (2023). Computational and artificial intelligence-based methods for antibody development. *Trends Pharmacol Sci.*, **44**(3), 175–189.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

L. Garner, A. and D. Janda, K. (2011). Protein-protein interactions and cancer: Targeting the central dogma. *Current Topics in Medicinal Chemistry*, **11**(3), 258–280.

Ledesma, F. G. and Mewes, M. (2020). Spherical-harmonic tensors. *Phys. Rev. Res.*, **2**, 043061.

Lensink, M. F. and Wodak, S. J. (2014). Score_set: A capri benchmark for scoring protein complexes. *Proteins: Structure, Function, and Bioinformatics*, **82**(11), 3163–3169.

Liao, Y.-L. and Smidt, T. (2023). Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *The Eleventh International Conference on Learning Representations*.

Liao, Y.-L., Wood, B., Das, A., and Smidt, T. (2023). Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. (2022). Evolutionary-scale prediction of atomic level protein structure with a language model.

Marks, C. and Deane, C. M. (2017). Antibody h3 structure prediction. *Computational and Structural Biotechnology Journal*, **15**, 222–231.

Martinkus, K., Ludwiczak, J., Cho, K., Liang, W.-C., Lafrance-Vanasse, J., Hotzel, I., Rajpal, A., Wu, Y., Bonneau, R., Gligorijevic, V., and Loukas, A. (2024). Abdiffuser: Full-atom generation of in vitro functioning antibodies.

McFee, M. and Kim, P. M. (2023). Gdockscore: a graph-based protein–protein docking scoring function. *Bioinformatics advances*, **3**(1).

Melnyk, I., Lozano, A., Das, P., and Chenthamarakshan, V. (2024). Alphafold distillation for protein design.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). Colabfold: making protein folding accessible to all. *Nature Methods*, **19**, 1–4.

Méndez, R., Leplae, R., De Maria, L., and Wodak, S. J. (2003). Assessment of blind predictions of protein-protein interactions: Current status of docking methods. *Proteins: Structure, Function, and Bioinformatics*, **52**(1), 51–67.

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., Dolma, S., Coulombe-Huntington, J., Chatr-aryamontri, A., Dolinski, K., and Tyers, M. (2020). The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, **30**(1), 187–200.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Renaud, N., Geng, C., Georgievska, S., Ambrosetti, F., Ridder, L., Marzella, D. F., Réau, M. F., Bonvin, A. M. J. J., and Xue, L. C. (2021). Deeprank: a deep learning framework for data mining 3d protein-protein interfaces. *Nature Communications*, **12**(1).

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, **12**, 77.

Roney, J. P. and Ovchinnikov, S. (2022). State-of-the-art estimation of protein model accuracy using alphafold. *Physical Review Letters*, **129**(23).

Réau, M., Renaud, N., Xue, L. C., and Alexandre (2022). Deeprank-gnn: a graph neural network framework to learn patterns in protein–protein interfaces. *Bioinformatics*, **39**(1).

Schneider, C., Buchanan, A., Taddese, B., and Deane, C. M. (2021). Dlab: deep learning methods for structure-based virtual screening of antibodies. *Bioinformatics*, **38**(2), 377–383.

Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A. A., and Wilson, A. G. (2021). Does knowledge distillation really work? *CoRR*, **abs/2106.05945**.

Steinegger, M. and Söding, J. (2017). Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, **35**(11), 1026–1028.

Stärk, H., Ganea, O.-E., Pattanaik, L., Barzilay, R., and Jaakkola, T. (2022). Equibind: Geometric deep learning for drug binding structure prediction.

Thomas, N., Smidt, T., Kearnes, S., Yang, L., Li, L., Kohlhoff, K., and Riley, P. (2018). Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds.

Tomezsko, P. J., Ford, C. T., Meyer, A. E., Michaleas, A. M., and Jaimes, R. (2024). Human cytokine and coronavirus nucleocapsid protein interactivity using large-scale virtual screens. *Frontiers in Bioinformatics*, **4**.

Tsuchiya, Y., Yamamori, Y., and Tomii, K. (2022). Protein–protein interaction prediction methods: from docking-based to ai-based approaches. *Biophysical Reviews*, **14**(6), 1341–1348.

Tyka, M. D., Keedy, D. A., André, I., DiMaio, F., Song, Y., Richardson, D. C., Richardson, J. S., and Baker, D. (2011). Alternate states of proteins revealed by detailed energy landscape mapping. *Journal of Molecular Biology*, **405**(2), 607–618.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, **abs/1706.03762**.

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., and 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.

Wallner, B. (2023). AFsample: improving multimer prediction with AlphaFold using massive sampling. *Bioinformatics*, **39**(9), btad573.

Wang, X., Terashi, G., Christoffer, C. W., Zhu, M., and Kihara, D. (2019). Protein docking model evaluation by 3d deep convolutional neural networks. *Bioinformatics*, **36**(7), 2113–2118.

Wang, X., Flannery, S. T., and Kihara, D. (2021). Protein docking model evaluation by graph neural networks. *Frontiers in Molecular Biosciences*, **8**.

Wang, Z., Combs, S. A., Brand, R., Calvo, M. R., Xu, P., Price, G., Golovach, N., Salawu, E. O., Wise, C. J., Ponnapalli, S. P., and Clark, P. M. (2022). Lm-gvp: an extensible sequence and structure informed deep learning framework for protein property prediction. *Scientific Reports*, **12**(1).

Xu, X. and Bonvin, A. M. J. J. (2024). Deeprank-gnn-esm: A graph neural network for scoring protein-protein models using protein language model. *Bioinformatics advances*.

Zhang, Y. (2008). Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, **18**(3), 342–348.

Zitnick, C. L., Das, A., Kolluru, A., Lan, J., Shuaibi, M., Sriram, A., Ulissi, Z. W., and Wood, B. M. (2022). Spherical channels for modeling atomic interactions. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*.