Preprint, 2024, 1–9
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: Matthew McFee matthew.mcfee@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. 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 a scoring functions. Beyond docking, scoring functions are a critical component of assessing structures produced by many protein generative models. Scoring models are also used as a final filtering in many generative deep learning models including those that generate antibody binders, and those which perform docking. In this work we present improved scoring functions for protein-protein interactions which utilizes cutting-edge euclidean graph neural network architectures, to assess protein-protein interfaces. These euclidean docking score models are known as EuDockScore, and EuDockScore-Ab with the latter being antibody-antigen dock specific. Finally, we provided EuDockScore-AFM a model trained on antibody-antigen outputs from AlphaFold-Multimer which proves useful in re-ranking large numbers of AlphaFold-Multimer outputs. 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 a 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 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 for structure prediction (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 process is typically broken down into two stages. The docking procedure 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

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

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.

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). Due to the limitations of these models, 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).

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 will be invariant.

1.3. Scoring functions in the era of AlphaFold

Generative deep learning models now dominate over traditional docking methods. For example, the bioinformatics field was revolutionized by the publication of AlphaFold2 (Jumper *et al.*, 2021), a deep learning model which can take in protein sequences and produce highly accurate structures. The AlphaFold2 architecture was then extended to generate the structures of multi-chain protein complexes in the form of AlphaFold-Multimer (AFM) (Evans *et al.*, 2022), providing an alternative avenue to computational docking techniques. Although AFM can produce accurate complex 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) and quality metrics (iPTM, plDDT, etc.) are still necessary to differentiate high-quality from low-quality models. Another example is AbDiffuser (Martinkus *et al.*, 2024), a deep learning model that generates antibody structures. It uses additional deep learning models that provide information such as predicted naturalness of predictions and predicted stability. Scoring models also remain a critical component of deep learning-based docking methods such as DiffDock (Corso *et al.*, 2023) and DiffDock-PP (Ketata *et al.*, 2023). Finally, DiffPack (Zhang *et al.*, 2024), a recent protein side-chain packing model, utilizes a confidence model. Thus, scoring models are still very useful supplemental tools in modern structural bioinformatics.

1.4. 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. Furthermore, antibody interface scoring is particularly 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 assisted by a specialized scoring function to help assess model generations.

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 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 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 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 further details.

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 that 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). 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 then used as input scalar features.

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. 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. Furthermore, we also compare our deep learning models to the more traditional machine learning scoring function, iScore (Geng *et al.*, 2019). 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). Additionally, we provide the maximum Matthews correlation coefficient (MCC) and F1 score across all thresholds. We found that in comparison to DeepRank and iScore, 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 ($p = 0.2468$). EuDockScore has the highest MCC and F1 score across all thresholds in comparison to all models. 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. The MCCs for GDockScore, EuDockScore, DeepRank, and iScore are 0.458. 0.538, 0.166, 0.279. The F1 scores for GDockScore, EuDockScore, DeepRank, and iScore are 0.528, 0.588, 0.270, 0.369. We further sought to investigate how the scores of near-native and incorrect

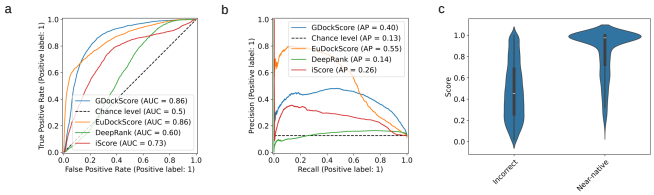

Fig. 1: (a) ROC curves comparing EuDockScore, GDockScore, DeepRank, and iScore. (b) Precision-recall curves of EuDockScore, GDockScore, Deeprank and iScore. (c) Violin plot comparing the distribution of scores produced by EuDockScore for the CAPRI score_set for near-native and incorrect decoys.

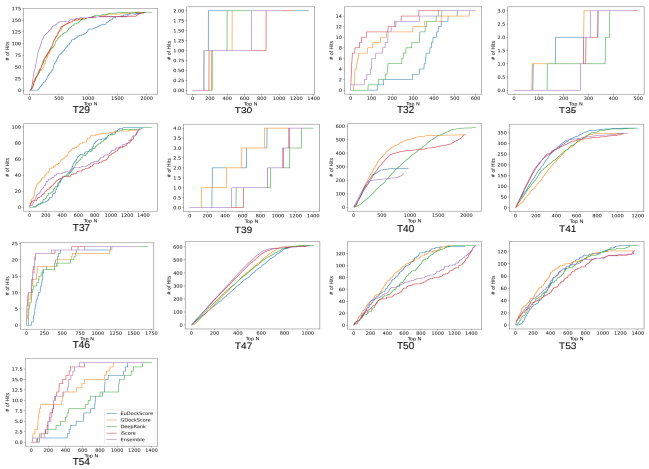

Fig. 2: Number of hits in selected models vs. top N selected curves for EuDockScore, GDockScore, DeepRank, iScore, and an ensemble average of all four models. Here a "hit" is defined as a near-native structure with a CAPRI quality of acceptable or better.

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 the 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.

Biologists only want to experimentally validate or study a handful of structure models they are confident are near-native, so a scoring model's ability to place 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 (Fig. 2) to be more comprehensive. In Fig. 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 but does worse on others. Ensembling improved overall ranking ability (Fig. 2) suggesting that multiple deep learning models contribute useful information to ranking proposed multimeric structural models.

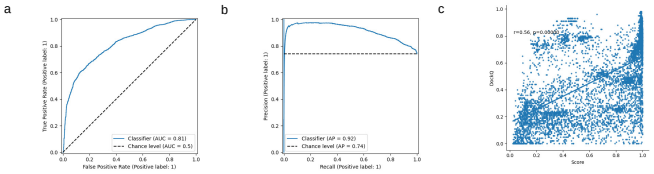

Fig. 3: (a) ROC curve for EuDockScore for dimers in the CASP15 competition. (b) Precision-recall for dimers in the CASP15 competition. (c) Scatter plot between EuDockScore scores and DockQ values. The r value and associated p-value are indicated on the plot.

2.4. EuDockScore performs well on CASP15 models generated by a variety of methods

To further test the capabilities of our model, we tested dimer models from CASP15 which included traditional modeling techniques and generative modeling including systems based on AlphaFold. We found that EuDockScore can differentiate high-quality vs. low-quality models for this set. We achieve an AUCROC of 0.81, an AP of 0.92 (Fig. 3 a,b), an MCC of 0.423, and an F1 of 0.868. Here we define a high-quality model as having a DockQ score >= 0.23. DockQ (Basu and Wallner, 2016) was selected as it combines many traditional model interface quality metrics into a single comprehensive metric. Furthermore, we find that our scores do correlate with DockQ scores and this correlation is statistically significant (Fig. 3c).

2.5. 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). Using a hold-out, non-redundant test set we see that EuDockScore-Ab can differentiate near-native from incorrect docks in unseen data (Fig. 4a,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. Furthermore, we found that the model can extrapolate to AlphaFold-Multimer generations but does not surpass the performance of AlphaFold-Multimer in our AUCROC or AP (Fig. 5a,b). This is likely due to the end-to-end nature of the AlphaFold-Multimer training protocol and overall model size and expressiveness.

2.6. EuDockScore-AFM: attempting to distil the scoring abilities of AFM type architectures with ground-truth correction

Since AlphaFold was previously used to rank monomeric structures (Roney and Ovchinnikov, 2022), GDockScore was successful in filtering AFM outputs (Tomezsko *et al.*, 2024) we next decided to distill AFM's scoring abilities into an independent model. Instead of engineering AlphaFold2 to output single chain like scores by utilizing decoys as templates such as AF2Rank (Roney and Ovchinnikov, 2022), we propose learning a new model.

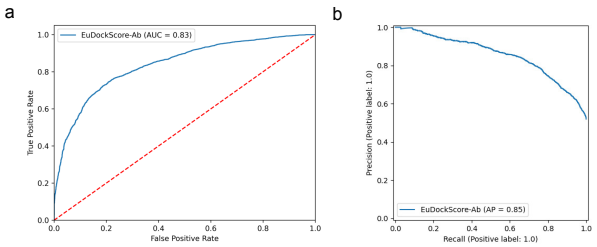

Fig. 4: (a) ROC curve for EuDockScore-Ab on hold-out, non-redundant antibody docks. (b) Precision-recall curve for EuDockScore-Ab on hold-out, non-redundant antibody docks.

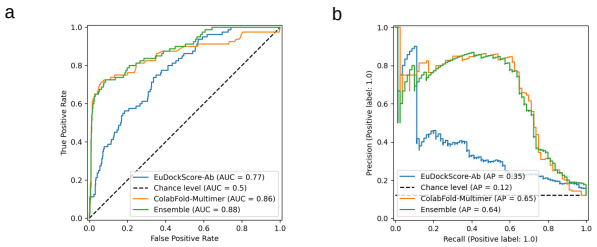

Fig. 5: (a) ROC curves comparing EuDockScore-Ab, ColabFold-Multimer (AFM ranking confidence (0.8*ipTM + 0.2*pTM)), and an ensemble of both scoring metrics (b) Precision-recall curves comparing EuDockScore-Ab, ColabFold-Multimer (AFM ranking confidence), and an ensemble of both scoring metrics.

Instead of simply training to distill AFM, the new training target is the sum of the AFM ranking confidence of the model and the DockQ against the native (true) structure. This combined metric allows our model to "correct" regions of the AFM energy function using ground truth structural data. When taking this new model, hereby known as EuDockScore-AFM, and testing it on a hold-out test set of non-redundant antibody structures generated by AFSample. It is important to note that we only use the heavy chain and antigen as inputs to our model as before. Although our model does possess ranking ability according to AUCROC and AP of the precision-recall curve (Fig. 6a,b), we do not outperform the stock AFM scoring metric or pDockQ2 (Zhu *et al.*, 2023), a recent scoring metric constructed from AFM quality predictions. Pairwise comparisons of all combinations of models result in statistically significant differences in AUCROC with $p \leq 2.2e-16$ for all comparisons. When we look at EuDockScore's ranking ability for individual DockQ quality classifications, there is a trend that higher quality structures have higher EuDockScore values (Fig. 6c). For AFSample, pDockQ2, EuDockScore, and the ensemble, the F1 scores are 0.631, 0.625, 0.324, and 0.587. The maximum MCCs achieved in the same order are 0.642, 0.645, 0.268, and 0.594.

More interestingly, when we actually compare the ranking abilities of each model for AFSample outputs, as in Figure 3. We find that there are many cases in which EuDockScore has superior ranking ability such as PDB IDs 1MLC, 5WK3, and 6A77 (Figs. 7,8). Ensembling all scoring systems also seems to help in many cases. This suggests that our model is useful in the context of helping re-rank model outputs from generative models such as AFM which has been shown in additional studies (Tomezsko *et al.*,

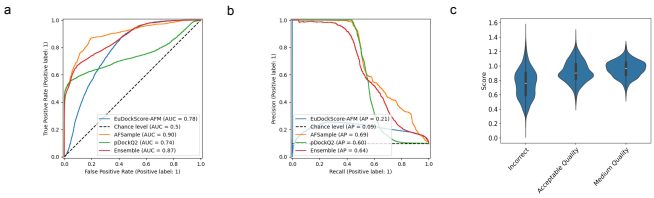

Fig. 6: (a) ROC curves comparing EuDockScore-AFM, AFSample (AFM ranking confidence (0.8*ipTM + 0.2*pTM)), pDockQ2, and an ensemble of both scoring metrics (b) Precision-recall curves comparing EuDockScore-Ab, AFSample (AFM ranking confidence), and an ensemble of both scoring metrics.

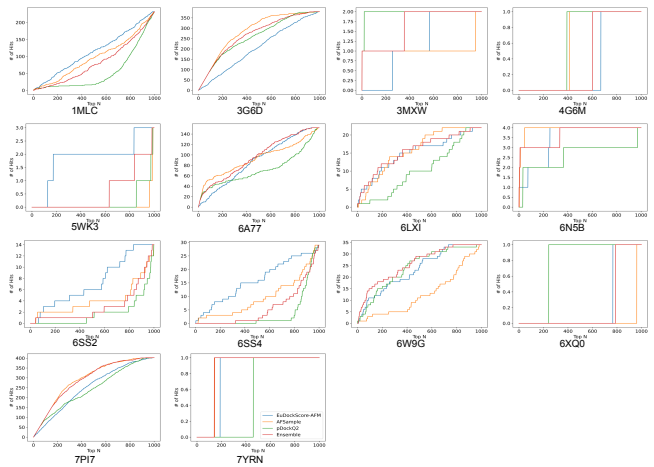

Fig. 7: Number of hits in selected top N models vs. top N selected models for EuDockScore-AFM, AFSample, pDockQ2, and an ensemble average of all four models. Here an acceptable model or "hit" has DockQ ≥ 0.23.

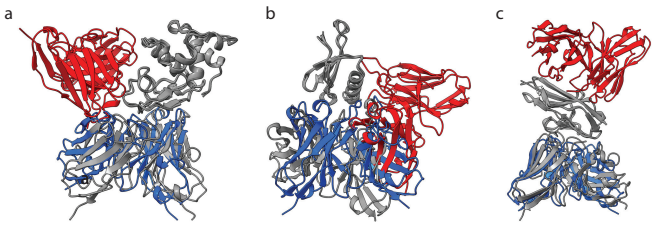

Fig. 8: Example structures of (A) 1MLC (B) 5WK3 and (C) 6A77. Structures are superimposed based on the antigen structure. Here, grey indicates the experimentally defined PDB structure. Blue indicates structures with good EuDockScore-AFM and DockQ, but poor AFM ranking confidence. Red indicates structures with poor EuDockScore-AFM and DockQ, but good AFM ranking confidence.

2024). Therefore, although current scoring models do not outright replace the quality metrics of generative models such as AFM, they do provide useful information for additional filtering of structure candidate models.

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), and iScore improves overall discriminative power suggesting the importance of using multiple deep learning models to assess the quality of generative modeling predictions. Our base model also generalizes well to models submitted to CASP15 further supporting the generalizability of our scoring system.

EuDockScore-Ab is a heavy chain-antigen specific scoring model trained on re-docked and relaxed real antibody complexes. EuDockScore-Ab can differentiate near-native (acceptable or better by traditional CAPRI metrics) from non-native antibody complexes. Additionally, EuDockScore-AFM a model trained specifically on AFM antibody-antigen complex predictions shows the improved ability to select out near-native structures in large pools of candidates that the standard AFM scoring metrics provided by ColabFold and AFSample showing the continued applicability of scoring function development. Although our scoring models prove useful, it is clear a significant degree of work must be done in the antibody space to achieve a state-of-the-art replacement for AlphaFold-Multimer metrics.

The work presented in this manuscript can be extended in many ways. For example, AFM models could be generated for the entire PDB to create an AFM-specific scoring function that isn't limited to antibodies. However, this will be a computationally intensive task. Recent literature has also suggested that transformers can be applied directly to atomistic data while achieving approximate equivariance. Transformers are significantly more scalable than current equivariant networks and may allow for improved learning.

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).

We used the CAPRI standard assessment for docking quality as specified in Mendez et al. (Méndez *et al.*, 2003). We assign non-native decoys as label 0 and acceptable or better decoys as

label 1. We generated roughly 50000 docking decoys across the AbDb with a 70%, 15%, and 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 unbiased 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 improve structure quality.

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 AFM and AFSample

To generate AFM predictions of Ab complexes, we used the protocol above, but instead generated models using the localcolabfold 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 were no cluster overlaps between PDBs used in model training and validation and those used for testing. To generate a test set with a large enough number of models for meaningful re-rankings, we used AFSample (Wallner, 2023) to generate a thousand models for each PDB assigned to the test set.

4.3. 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).

4.4. 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.

4.5. Model implementation, hyperparameter tuning, and training details

Hyperparameter tuning was performed by adjusting the number of spherical channels ie. the 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). The cross-entropy loss was used for the loss function during training with binary labels as specified in (McFee and Kim, 2023) and the mean-squared-error loss for training on continuous targets as in EuDockScore-AFM.

4.6. Definition of average precision

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

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

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.

4.7. Data analysis and statistical Testing

The ROCs, precision-recall curves, Matthews correlation coefficients, and F1 scores in this paper were computed 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 and JK report no conflicts.

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 by PMK.

7. Acknowledgments

We would like to thank the Digital Research Alliance of Canada for providing computing resources. CIHR Project Grants PJT-159750 and PJT-166008 funded 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. 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 CASP15 data, including DockQ values, was download from `https://predictioncenter.org/download_area/`. 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.

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.

Basu, S. and Wallner, B. (2016). Dockq: A quality measure for protein-protein docking models. *PLOS ONE*, **11**(8), e0161879.

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.

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.

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.

Geng, C., Jung, Y., Renaud, N., Honavar, V., Bonvin, A. M. J. J., and Xue, L. C. (2019). iScore: a novel graph kernel-based function for scoring protein–protein docking models. *Bioinformatics*, **36**(1), 112–121.

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.

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).

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.

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.

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.

Zhang, Y., Zhang, Z., Zhong, B., Misra, S., and Tang, J. (2024). Diffpack: A torsional diffusion model for autoregressive protein side-chain packing.

Zhu, W., Shenoy, A., Kundrotas, P. J., and Elofsson, A. (2023). Evaluation of alphafold-multimer prediction on multi-chain protein complexes. *Bioinformatics*, **39**(7).

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*.