Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

A Gated Graph Transformer for Protein Complex Structure Quality Assessment and its Performance in CASP15

Xiao Chen [1,+], Alex Morehead [1,+], Jian Liu [1,], and Jianlin Cheng [1,*]

[1] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, 65201, MO, USA

+: Equal Contribution. *To whom correspondence should be addressed.

Abstract

Motivation: Proteins interact to form complexes to carry out essential biological functions. Computational methods such as AlphaFold-multimer have been developed to predict the quaternary structures of protein complexes. An important yet largely unsolved challenge in protein complex structure prediction is to accurately estimate the quality of predicted protein complex structures without any knowledge of the corresponding native structures. Such estimations can then be used to select high-quality predicted complex structures to facilitate biomedical research such as protein function analysis and drug discovery.
Results: In this work, we introduce a new gated neighborhood-modulating graph transformer to predict the quality of 3D protein complex structures. It incorporates node and edge gates within a graph transformer framework to control information flow during graph message passing. We trained, evaluated and tested the method (called DProQA) on newly-curated protein complex datasets before the 15th Critical Assessment of Techniques for Protein Structure Prediction (CASP15) and then blindly tested it in the 2022 CASP15 experiment. The method was ranked 3rd among the single-model quality assessment methods in CASP15 in terms of the ranking loss of TM-score on 36 complex targets. The rigorous internal and external experiments demonstrate that DProQA is effective in ranking protein complex structures.
Availability: The source code, data, and pre-trained models are available at https://github.com/jianlin-cheng/DProQA
Contact: chengji@missouri.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Proteins perform a broad range of biological functions. Protein-protein interactions (PPI) play a key role in many biological processes. Understanding the mechanisms and functions of PPI may benefit many areas such as drug discovery (Scott *et al.* (2016); Athanasios *et al.* (2017); Macalino *et al.* (2018) and protein design Kortemme and Baker (2004); Baker (2006); Lippow and Tidor (2007)). Typically, high-resolution 3D structures of protein complexes can be determined using experimental solutions (e.g., X-ray crystallography and cryo-electron microscopy). However, due to the high costs associated with them, these methods cannot meet all the increasing demand of protein complex structures in modern biological research and technology development. In the context of this practical challenge, computational methods for protein complex structure prediction have recently been receiving an increasing amount of attention.

Recently, AlphaFold2-Multimer (Evans *et al.* (2021)), an end-to-end system for protein complex structure prediction system, improves the accuracy of predicting multimer (quaternary) structures considerably. However, compared to AlphaFold2's outstanding performance for monomer (tertiary) structure prediction (Jumper *et al.* (2021)), the accuracy level for protein quaternary structure prediction still has much room for progress. One specific problem of predicting quaternary structures is the estimation of model accuracy (EMA) (also called quality assessment(QA), which plays a significant role in ranking and selecting predicted quaternary structural models of good quality (Kinch *et al.* (2021)).

However, unlike the tertiary structure quality assessment with many machine learning, particularly deep learning methods developed for evaluating the quality of tertiary structural models over many years, there are few deep learning methods for evaluating the quality of quaternary structural models. Existing EMA methods have not leveraged

cutting-edge attention-based transformer-like deep learning architectures (Vaswani *et al.* (2017)) to enhance the expressiveness of structural quality assessment. The models in most current training or benchmark datasets for quaternary structure quality assessment (Liu *et al.* (2008); Lensink and Wodak (2014); Kotthoff *et al.* (2021)) were generated by traditional protein docking methods (Tovchigrechko and Vakser (2006); Pierce *et al.* (2011)), which do not effectively represent the quaternary structural models of generally better quality predicted by the latest protein complex structure predictors (Jumper *et al.* (2021); Bryant *et al.* (2022)). Consequently, EMA methods trained using these datasets may not work well on the structural models generated by the latest, more accurate protein complex structure predictors.

In general, EMA methods for tertiary structures can be divided into two categories: multi-model methods and single-model methods. Multi-model methods take a pool of protein structural models as input and may use a comparison between the models to evaluate their quality, such as in the procedure performed by Pcons (Lundström *et al.* (2001)), ModFOLDClust (McGuffin (2007)), and DeepRank2 (Chen *et al.* (2021)). In contrast, single-model methods give a certain quality score for each protein structural model without considering other models' information, such as in the procedure performed by ProQ2 (Uziela and Wallner (2016)), ProQ3 (Uziela *et al.* (2016)), DISTEMA (Chen and Cheng (2022)), and GNN_DOVE (Wang *et al.* (2021)). In this work, we develop a single-model Deep Protein Quality Assessment method (DProQA) for predicting the quality of protein complex structural models. In particular, DProQA introduces a Gated Graph Transformer, a novel graph neural network (GNN) that learns to modulate its input information to better guide its structure quality predictions. DProQA takes a single protein complex structure as input to predict its quality via a single forward pass and therefore is pretty fast. Moreover, it uses a multi-task learning strategy to predict the real-valued quality score of a structural model as well as classify it into multiple quality categories.

2 Related Work

Protein tertiary and quaternary structure prediction. Predicting protein structures has been an essential problem for the last several decades. Recently, the problem of protein tertiary structure prediction has largely been solved by deep learning methods (e.g., Jumper *et al.* (2021)). Furthermore, new deep learning methods (e.g., Evans *et al.* (2021); Bryant *et al.* (2022); Guo *et al.* (2022)) have begun making advancements in protein complex (quaternary) structure prediction.

Protein representation learning. Protein structures can be represented in various ways. Previously, proteins have been represented as tableau data in the form of hand-crafted features (Chen *et al.* (2020)). Along this line, many works (Wu *et al.* (2021); Chen and Cheng (2022)) have represented proteins using pairwise information embeddings such as residue-residue distance maps and contact maps. Recently, describing proteins as graphs has become a popular means of representing proteins, as such representations can learn and leverage proteins' geometric information more naturally. For example, EnQA (Chen *et al.* (2023)) used 3D-equivariant graph representations to estimate the per-residue quality of protein structures.

Machine learning for protein structure quality assessment. Over the past few decades, various EMA methods for ranking and scoring protein complex structural models have been developed (Gray *et al.* (2003); Huang and Zou (2008); Vreven *et al.* (2011); Basu and Wallner (2016b); Geng *et al.* (2020)). Among these scoring methods, machine learning-based EMA methods have shown better performance than physics-based (Dominguez *et al.* (2003); Moal *et al.* (2013) and statistics-based methods Zhou and Zhou (2002); Pons *et al.* (2011)).

Recent machine learning methods have utilized various techniques and features to approach the task. For example, ProQDock (Basu and Wallner (2016b)) and iScore (Geng *et al.* (2020)) used protein structural features as the input for a support vector machine to predict model quality. EGCN (Cao and Shen (2020)) assembled graph pairs to represent protein complex structures and then employed a graph convolutional network (GCN) to learn graph structural information. DOVE (Wang *et al.* (2020)) used a 3D convolutional neural network (CNN) to extract features from protein-protein interfaces to predict model quality. In a similar spirit, GNN_DOVE (Wang *et al.* (2021)) and PPDocking (Han *et al.* (2021)) trained Graph Attention Networks (Veličković *et al.* (2017)) to evaluate protein complex decoys. Moreover, PAUL (Eismann *et al.* (2021)) used a rotation-equivariant neural network to identify accurate models of protein complexes.

Deep transformers. Increasingly more works have applied transformer-like architectures or multi-head attention (MHA) mechanisms to achieve state-of-the-art results in different domains. For example, the Swin-Transformer achieved state-of-the-art performance in various computer vision tasks (Hu *et al.* (2019); Liu *et al.* (2021, 2022)). Likewise, the MSA Transformer (Rao *et al.* (2021)) used tied row and column MHA to extract features from multiple sequence alignments of proteins. Moreover, DeepInteract (Morehead *et al.* (2021b)) introduced the Geometric Transformer to model protein chains as graphs for protein interface contact prediction.

Contributions. Our work builds upon prior works by making the following contributions:

1. We provide the *first* example of applying transformer representation learning to the task of protein complex structure quality assessment, by introducing the new gated graph transformer architecture to iteratively update node and edge representations using the adaptive feature modulation.
2. The DProQA method was trained using the newly-developed protein complex datasets in which all structural decoys were generated using AlphaFold2 (Jumper *et al.* (2021)) and AlphaFold-Multimer (Evans *et al.* (2021)).
3. Using the newly-developed Docking Benchmark 5.5-AF2 (DMB55-AF2), we demonstrate the state-of-the-art performance of DProQA in comparison with the existing methods such as ZRANK2 (Pierce and Weng (2008)), GOAP (Zhou and Skolnick (2011)) and GNN_DOVE (Wang *et al.* (2021)).
4. DProQA was blindly tested in the 15th community-wide Critical Assessment of protein Structure Prediction (CASP15) in 2022, where it ranked 3rd among all single-model EMA methods in terms of ranking loss of structural models' TM-score.

3 Methods and Materials

Illustrated from left to right in Figure 1, DProQA first receives a 3D protein complex structure as input and represents it as a K-NN graph. Notably, all chains in the complex are represented within the same graph, where pairs of atoms from the same chain are distinguished using a binary edge feature (i.e., in the same chain or not). Therefore, it can uniformly deal with a complex consisting of any number of chains. Moreover, it only requires a single protein complex structure as input without using any extra information such as multiple sequence alignments (MSAs) and residue-residue co-evolutionary features extracted from MSAs. Its output includes a real-valued quality score of the structure as well as a quality class it is assigned to.

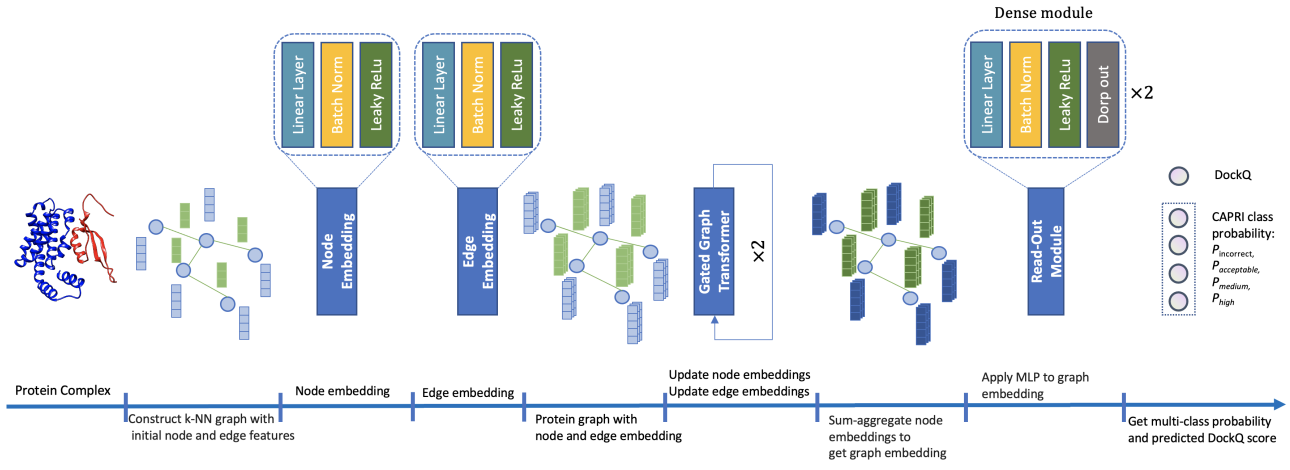

Fig. 1. An overview of the DProQA pipeline for quality assessment of protein complexes. The input is a protein complex structure. The output includes the predicted quality score of the input structure (e.g., DockQ score) and the probability of the quality class (i.e., incorrect topology, acceptable quality, medium quality, and high quality) which the structure is classified into.

3.1 K-NN graph representation of protein complex structure

K-NN graph representation K-nearest neighbors (K-NN) has been used in many previous studies for protein structure analysis and molecular representation learning (Ingraham *et al.* (2019)). Moreover, using other graph construction techniques (e.g. distance-based definitions of edges) may introduce inherent graph-structural biases (Jing *et al.* (2020)) within a network's learning process. Subsequently, in this work, DProQA represents each input protein complex structure as a spatial k-nearest neighbors (k-NN) graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the protein's C$\alpha$ atoms serve as $\mathcal{V}$ (i.e., the nodes of $\mathcal{G}$). After constructing $\mathcal{G}$ by connecting each node to its 10 closest neighbors in $\mathbb{R}^3$, we denote its initial node features as $\mathbf{F}$ (e.g., residue type) and its initial edge features as $\mathbf{E}$ (e.g., Cα-Cα distance), as described more thoroughly in Supplementary Section Featurization and Table S1.

Node and edge embeddings. After receiving a protein complex graph $\mathcal{G}$ as input, DProQA applies initial node and edge embedding modules to each node and edge, respectively. We define such embedding modules as φ_N and φ_E respectively, where each φ function is represented as a shallow neural network consisting of a linear layer, batch normalization, and LeakyReLU activation function (Xu *et al.* (2015)). Such node and edge embeddings are then fed as an updated input graph to the Gated Graph Transformer.

3.2 Gated graph transformer architecture

Unlike other graph neural network (GNN)-based structure scoring methods (Wang *et al.* (2021); Han *et al.* (2021)), which define edges using a fixed distance threshold so that each graph node may have a different number of incoming and outgoing edges, DProQA constructs and operates on k-NN graphs where all nodes are connected to the same number of neighbors. However, in the context of k-NN graphs, each neighbor's information is, by default, given equal priority during information updates. Here, we may desire to imbue our graph neural network with the ability to automatically decide the priority of different nodes and edges during the graph message passing. Consequently, we design GGT, a gated neighborhood-modulating graph transformer inspired by Veličković *et al.* (2017); Dwivedi and Bresson (2020); Morehead *et al.* (2021b) to update the features of the nodes and edges. Formally, to update the network's node embeddings $\mathbf{h}_i$ and edge embeddings $\mathbf{e}_{ij}$, we define a single layer of the GGT as:

$$\widehat{\widehat{\mathbf{w}}}_{ij}^{k,\ell} = \left(\frac{Q^{k,\ell}\mathbf{h}_i^{\ell} \cdot K^{k,\ell}\mathbf{h}_j^{\ell}}{\sqrt{d_k}} \right) \cdot E^{k,\ell}\mathbf{e}_{ij}^{\ell} \tag{1}$$

$$\hat{\mathbf{e}}_{ij}^{\ell+1} = \mathbf{O}_e^{\ell} \|_{k=1}^{H} \left(\widehat{\widehat{\mathbf{w}}}_{ij}^{k,\ell} \right) \tag{2}$$

$$\widehat{\mathbf{w}}_{ij}^{k,\ell} = \widehat{\widehat{\mathbf{w}}}_{ij}^{k,\ell} \times \text{sigmoid} \left(\mathbf{G}_e^{k,\ell}\mathbf{e}_{ij}^{\ell} \right) \tag{3}$$

$$\mathbf{w}_{ij}^{k,\ell} = \text{softmax} \left(\hat{\mathbf{w}}_{ij}^{k,\ell} \right) \tag{4}$$

$$\hat{\mathbf{h}}_i^{\ell+1} = O_{\mathbf{h}}^{\ell} \|_{k=1}^{H} \left(\text{sigmoid} \left(G_h^{k,\ell}\mathbf{h}_j^{\ell} \right) \times \sum_{j \in \mathcal{N}_i} \mathbf{w}_{ij}^{k,\ell} V^{k,\ell}\mathbf{h}_j^{\ell} \right) \tag{5}$$

In particular, the GGT adds on top of the standard graph transformer architecture (Dwivedi and Bresson (2020)) two information gates through which the network can modulate node and edge information flow, as shown in Figure 2. Several main operations in Figure 2 are described by the following equations. Equation 1 calculates the attention score with additional edge information by multiplying edge embedding $\mathbf{e}_{ij}^{\ell}$. Equation 2 is used to update edge embeddings. Here $\mathbf{O}_e^{\ell}$ denotes an edge representation update multi-layer perceptron (MLP); $||$ denotes concatenation; and H is the number of attention heads. Equation 3 shows how the output of the edge gate, $\mathbf{G}_e$, is used element-wisely to gate the network's attention scores $\widehat{\widehat{\mathbf{w}}}_{ij}$ derived from Equation 1. Such gating simultaneously allows the network to regulate the edge information flow. Equation 4 converts an attention score $\hat{\mathbf{w}}_{ij}^{k,\ell}$ to a probability score $\mathbf{w}_{ij}^{k,\ell}$ by softmax function. The node gate, $\mathbf{G}_h$ in Equation 5 allows the GGT to individually control the information flow from each node neighbor during the node information update. Lastly, for the GGT's feed-forward network (FFN), we use the same structure as described in Dwivedi and Bresson (2020), providing us with the new node embeddings $\hat{\mathbf{h}}_i^L$ and edge embeddings $\hat{\mathbf{e}}_{ij}^L$.

3.3 Multi-task graph property prediction

To obtain graph-level predictions for each input protein complex graph, we apply a graph sum-pooling operator on $\hat{\mathbf{h}}_i^L$ to get the graph embedding

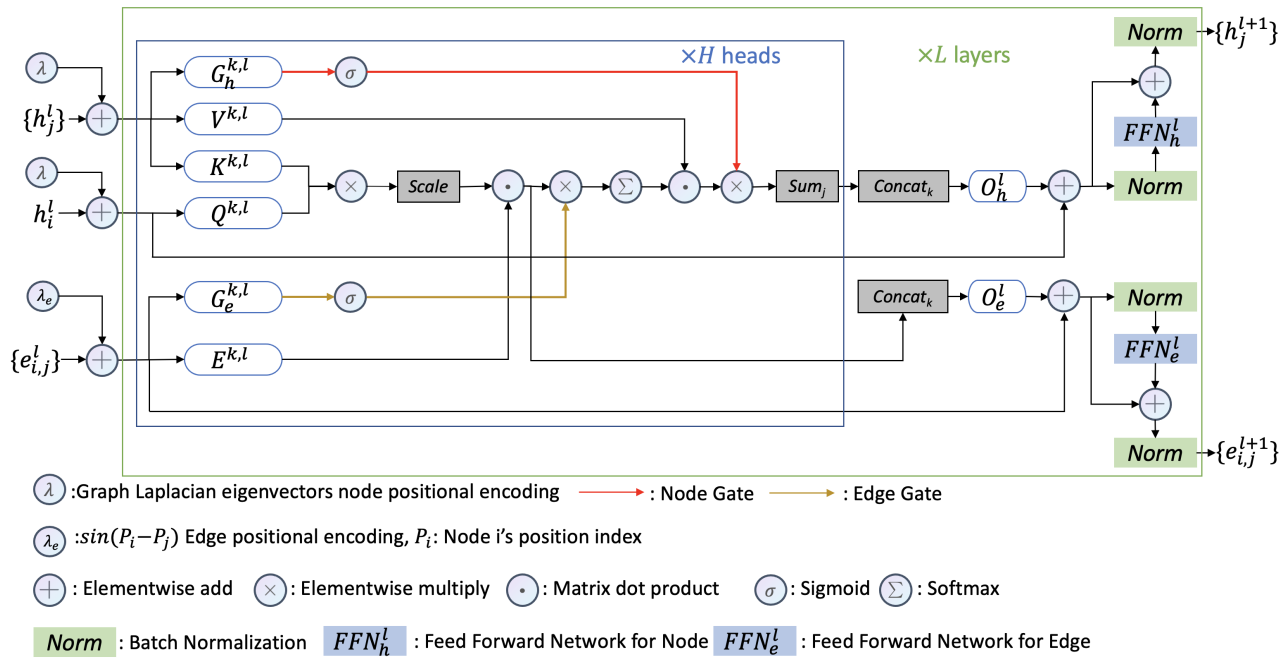

Fig. 2. The gated graph transformer (GGT) model architecture for updating the features of nodes of a protein complex graph.

$\mathbf{p}$. This graph embedding $\mathbf{p}$ is then fed as input to DProQA's two read-out modules, where each read-out module consists of a series of linear layers, batch normalization layers, LeakyReLU activation, and dropout layers (Hinton *et al.* (2012)), respectively. The output from the read-out modules is used by a softmax function in the classification output layer to classify the input into the four different quality classes (i.e., Incorrect, Acceptable, Medium, and High). The output from the read-out modules is also used by the regression output layer with a sigmoid activation function to $\mathbf{y}$ to obtain a single scalar output representing the predicted DockQ score (Basu and Wallner (2016a)) for the protein complex input.

Structural quality score prediction loss. To train DProQA's graph regression head, we used the mean squared error loss $\mathcal{L}_R = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{q}'_i - \mathbf{q}^*_i\|^2$. Here, $\mathbf{q}'_i$ is the model's predicted DockQ score for example i, $\mathbf{q}^*_i$ is the ground truth DockQ score for example i, and N represents the number of examples in a given mini-batch.

Structural quality classification loss. To train DProQA's graph classification head, we used the cross-entropy loss $\mathcal{L}_C = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{C}\sum_{j=1}^{C}\left(\mathbf{y}^*_{ij}\cdot\log(\mathbf{y}'_{ij})\right)$. Here, $\mathbf{y}'_{ij}$ is the predicted probability of the model's quality belonging to class j for example i, and $\mathbf{y}^*_{ij}$ is the ground truth DockQ quality class j (e.g., Incorrect) for example i. N denotes the number of examples in a given mini-batch and C for the number of classes.

Overall loss. DProQA's overall loss is the weighted sum of the two losses above: $\mathcal{L} = w_{\mathcal{L}_C}(0.1)\times\mathcal{L}_C + w_{\mathcal{L}_R}(0.9)\times\mathcal{L}_R$. The weights for each constituent loss (e.g., $w_{\mathcal{L}_R}$) were determined either by performing a grid search or using a lowest-validation-loss criterion for parameter selection.

Supplementary Section Implementation Details and Table S2 describe how we implemented, trained, and tuned the DProQA.

3.4 Training and test data

Multimer-AF2 training dataset. Similar to Morehead *et al.* (2021a), we created a new Multimer-AF2 (MAF2) dataset comprised of multimeric structures predicted by AlphaFold 2 (Jumper *et al.* (2021)) and AlphaFold-Multimer (Evans *et al.* (2021)) structure prediction pipeline on the Summit supercomputer (Gao *et al.* (2021, 2022)). The protein multimer targets for which we predicted structures were obtained from the EVCoupling (Hopf *et al.* (2019)) and DeepHomo (Yan and Huang (2021)) datasets, which consist of both heteromers and homomers. In summary, the MAF2 dataset contains a total of $9,251$ decoys. According to DockQ scores, 20.44% of them are of Incorrect quality, 14.34% of them are of Acceptable quality, 30.00% of them are of Medium quality, and the remaining 35.22% of them are of High quality.

Docking Decoy Set for training. The Docking Decoy dataset (Kundrotas *et al.* (2018)), contains 58 protein complex targets. Each target includes approximately 100 incorrect decoys and at least one *near-native* decoy.

The Docking Decoy set and MAF2 set were used together as the training data to train and validate the DProQA, which together include $12,040$ decoys in total. To split the data into training and validation sets, we applied MMseq2 (Mirdita *et al.* (2021)) to cluster all targets' sequences with 30% sequence identity. Then we selected 70% of the clusters' decoys as the training set and the rest as the validation set. The combined training set contains $8,733$ decoys, and the validation set contains $3,407$ decoys.

Docking Benchmark5.5 AF2 test dataset. The Docking Benchmark 5.5 AF2 (DBM55-AF2) dataset is the first test dataset. We applied AlphaFold-Multimer (Evans *et al.* (2021)) to predict the structures of Docking Benchmark 5.5 targets (Vreven *et al.* (2015)). This test dataset contains a total of 15 protein targets with 449 total decoy models, 50.78% of these decoys are of Incorrect quality, 16.70% of them are of Acceptable quality, 30.73% of them are of Medium quality, and the remaining 1.78% of them are of High quality.

More details about MAF2 and DBM55-AF2 generation and how we conducted sequence filtering and selected targets as the blind test set can be found in Supplementary Section Addition Dataset Information.

CASP15 EMA experiment. We blindly tested DProQA (Group name: MULTICOM_egnn, ID: 120) in 2022 CASP15 EMA category. DProQA was evaluated on 36 protein complex targets whose exprimental structures were available for us. Each target contains around 350 models from different CASP15 protein quaternary structure predictors.

Training labels. DProQA performs two learning tasks simultaneously. In the regression task, DProQA treats true DockQ scores (Basu and Wallner (2016a)) of decoys as its labels. As introduced earlier, DockQ scores are continuous values in the range of [0, 1]. A higher DockQ score indicates a higher-quality structure. In the classification task, DProQA predicts the probabilities that the structure of an input protein complex falls into the Incorrect, Acceptable, Medium, or High-quality category. Labeling a decoy into such quality categories was made according to its true DockQ score.

The true DockQ scores of the models in MAF2 and DBM55-AF2 were calculated by using the DockQ tool (Basu and Wallner (2016a)) to compare them with their corresponding true structures. The Docking Decoy Set provides interface root mean squared deviations (iRMSDs), ligand RMSDs (LRMSs), and fractions of native contacts (f_{nat}) for each decoy. We used Equation 1 and 2 in Basu and Wallner (2016a) to convert these scores to DockQ scores. The DockQ scores were then converted into four discrete categories: Incorrect, Acceptable, Medium, or High quality according to Basu and Wallner (2016a).

3.5 Evaluation setting

Baseline methods. We compared DProQA with three typical methods: ZRANK2, GOAP and GNN_DOVE. ZRANK2 is a method using a linear weight scoring function for evaluating protein complex structures. GOAP score is composed of all-atoms level distance-dependent and orientation-dependent potentials. GNN DOVE is an atom-level graph attention-based method for protein complex structure evaluation. It extracts the interface areas of a protein complex structure to build its input graph.

DProQA variants. Besides the standard DProQA model, we also report results on the DBM55-AF2 dataset for a selection of DProQA variants curated in this study. The DProQA variants includes DProQA_GT which employs the original Graph Transformer architecture Dwivedi and Bresson (2020); DProQA_GTE which employs the GGT with only its edge gate enabled; and DProQA_GTN which employs the GGT with only its node gate enabled.

Evaluation metrics. We evaluated the methods using two main metrics. The first metric measures how many qualified decoys are found within a model's predicted Top-N structure ranking for a target. Within this framework, a method's overall hit (success) rate is defined as the number of protein complex targets for which it ranks at least one acceptable, medium or higher-quality decoy within its Top-N ranked decoys, which is a metric used by the Critical Assessment of Protein-Protein Interaction (CAPRI) (Lensink and Wodak (2014)). In this work, we report the methods' Top-10 hit rates. A hit rate is represented by three numbers separated by the character /. These three numbers, in order, represent how many decoys with Acceptable or higher-quality, Medium or higher-quality, and High quality are among the Top-N ranked decoys. The second metric measures the ranking loss for each method. Here, the per-target ranking loss is defined as the difference between the DockQ score of a target's best decoy and the DockQ score of the top decoy selected by the ranking method. As such, a lower ranking loss indicates a better ranking ability.

4 Results

4.1 Performance on the DBM55-AF2 dataset

Table 1 presents the ranking loss for all methods on the DBM55-AF2 dataset. DProQA achieves the best ranking loss of 0.049 which is 83.6% lower than ZRANK2's ranking loss of 0.299, 60.2% lower than GOAP's ranking loss of 0.123 and 87% lower than GNN_DOVE's ranking loss of 0.379. Furthermore, for 4 targets, DProQA correctly selects the Top-1 model and achieves 0 ranking loss. Additionally, DProQA and GOAP achieve the lowest loss on 6 targets, while ZRANK2 gets the lowest loss on 4 targets and GNN_DOVE on 1 target. Notably, DProQA_GT, DProQA_GTE, and DProQA_GTN's losses are also lower than the three baseline methods.

Table 2 summarizes all the methods' hit rates on the DBM55-AF2 dataset, which contains 15 targets. DProQA achieves the best hit rate for ranking medium- and high-quality decoys together with ZRANK2, successfully selecting some medium- and high-quality decoys for all 10 targets and all 3 targets that have such models, respectively. However, it is ranked behind ZRANK2 and GOAP in selecting acceptable-quality decoys, i.e., it was able to rank at least one acceptable quality model in the top 10 for 12 targets, one fewer than ZRNK2 and GOAP, but higher than 10 of GNN_DOVE.

4.2 Impact of node and edge gates

The results in Tables 1 and 2 show that using both edge and node gates (DProQA) performs better than using only the edge gates (DProQA_GTE) or the node gates (DProQA_GTN), whose accuracy is better than or equal to not using any gate (DProQA_GT). Therefore, this ablation study specifically demonstrates that the edge and node gates are useful for predicting the quality of protein complex structures.

4.3 Performance in 2022 CASP15 EMA experiment

DProQA (team: MULTICOM_egnn) participated in the Estimation Model Accuracy(EMA) category of the CASP15 running from May to August 2022. We collected all EMA prediction results from CASP15 website CASP15 (2022) and used US-align Zhang *et al.* (2022) to calculate the TM-score Zhang and Skolnick (2004) ranking loss for all 36 targets whose true structures were available for us to perform evaluation. Figure 3 reports all CASP15 single-model EMA methods' average TM-score ranking loss, where DProQA ranked **3rd**. DProQA achieved a 0.200 ranking loss. All single-model methods' average ranking loss is 0.307.

Figure 4 illustrates the distribution of the MULTICOM_egnn's ranking loss on the 36 CASP15 EMA targets. The vertical dashed black line is the mark for the mean value. 23 out of 36 data points are located on the left side of the black line. This loss distribution is right-skewed, where the skewness value is 0.751.

Figure 5 shows that MULTICOM_egnn successfully selected a high-quality model with a very low ranking loss (i.e., 0.0014) for target H1111 (PDB code: 7QIJ) which is a Hetero 27-mer with a sequence length of 940. Figure 5(a) is the histogram of all server methods' model TM-scores for target H1111. Most models' TM-scores are low, yet still some are high-quality prediction models. In Figure 5(b), from left to right, the three protein complex structures shown are the corresponding native structure, the true TOP-1 model, and the MULTICOM_egnn top selected model, respectively. The top model selected by MULTICOM_egnn has a high TM-score of 0.9816. The ranking loss of MULTICOM_egnn for this target is the lowest among all the CASP15 EMA methods.

Table 1. The ranking loss on the DBM55-AF2 dataset. The overall row represents the mean and standard deviation of the ranking losses for all targets.

Target	DProQA	DProQA_GT	DProQA_GTE	DProQA_GTN	GOAP	ZRANK2	GNN_DOVE
6AL0	0.0	0.156	0.156	0.0	0.331	0.016	0.424
3SE8	0.079	0.041	0.041	0.079	0.0	0.677	0.735
5GRJ	0.024	0.012	0.095	0.012	0.23	0.757	0.776
6A77	0.037	0.062	0.0	0.037	0.59	0.539	0.591
4M5Z	0.015	0.026	0.026	0.015	0.133	0.221	0.221
4ETQ	0.0	0.76	0.0	0.748	0.0	0.759	0.759
5CBA	0.052	0.038	0.052	0.058	0.007	0.047	0.019
5WK3	0.114	0.114	0.114	0.186	0.109	0.0	0.087
5Y9J	0.0	0.0	0.0	0.0	0.0	0.027	0.382
6BOS	0.081	0.081	0.0	0.0	0.09	0.006	0.081
5HGG	0.051	0.051	0.121	0.051	0.051	0.051	0.121
6A0Z	0.207	0.207	0.207	0.207	0.214	0.012	0.062
3U7Y	0.0	0.021	0.0	0.0	0.0	0.772	0.756
3WD5	0.011	0.011	0.011	0.0	0.011	0.595	0.672
5KOV	0.065	0.08	0.085	0.087	0.078	0.01	0.0
Summary	**0.049 ± 0.054**	0.111 ± 0.182	0.061 ± 0.064	0.099 ± 0.185	0.123 ± 0.025	0.299 ± 0.104	0.379 ± 0.298

Table 2. Per-target and overall hit rates on the DBM55-AF2 dataset. The BEST column represents each target's best-possible Top-10 result, which is an upper limit of the hit rates. The Summary row lists the overall results of a method on all the targets.

Target	DProQA	DProQA_GT	DProQA_GTE	DProQA_GTN	GOAP	ZRANK2	GNN_DOVE	BEST
6AL0	9/2/0	10/0/0	10/0/0	10/2/0	6/0/0	9/0/0	9/0/0	10/2/0
3SE8	8/8/0	9/9/0	8/8/0	8/8/0	3/0/0	8/8/0	2/2/0	10/10/0
5GRJ	10/10/0	9/9/0	10/10/0	9/9/0	3/2/0	10/9/0	5/4/0	10/10/0
6A77	7/7/0	7/7/0	8/8/0	8/8/0	0/0/0	3/3/0	4/4/0	8/8/0
4M5Z	10/10/1	10/10/0	10/10/0	10/10/0	10/10/0	10/10/1	10/10/1	10/10/1
4ETQ	1/1/0	1/1/0	1/1/0	1/1/0	0/0/0	1/1/0	1/1/0	1/1/0
5CBA	10/10/1	10/10/0	10/10/0	10/10/1	10/10/3	10/10/2	10/10/4	10/10/6
5WK3	0/0/0	0/0/0	0/0/0	0/0/0	1/0/0	0/0/0	3/0/0	3/0/0
5Y9J	4/0/0	6/0/0	5/0/0	4/0/0	0/0/0	5/0/0	5/0/0	8/0/0
6BOS	10/10/0	10/10/0	10/10/0	10/10/0	10/10/0	10/10/0	10/10/0	10/10/0
5HGG	8/0/0	8/0/0	8/0/0	8/0/0	8/0/0	10/0/0	10/0/0	10/0/0
6A0Z	0/0/0	0/0/0	0/0/0	0/0/0	2/0/0	0/0/0	0/0/0	3/0/0
3U7Y	2/2/1	2/2/1	2/2/1	2/1/0	2/2/1	2/2/1	1/1/1	2/2/1
3WD5	10/8/0	9/8/0	9/8/0	9/8/0	0/0/0	10/8/0	6/4/0	10/10/0
5KOV	0/0/0	0/0/0	0/0/0	0/0/0	1/0/0	1/0/0	0/0/0	2/0/0
Summary	12/**10/3**	12/9/1	12/9/1	12/10/1	**13**/9/**3**	**13/10/3**	10/4/1	15/10/3

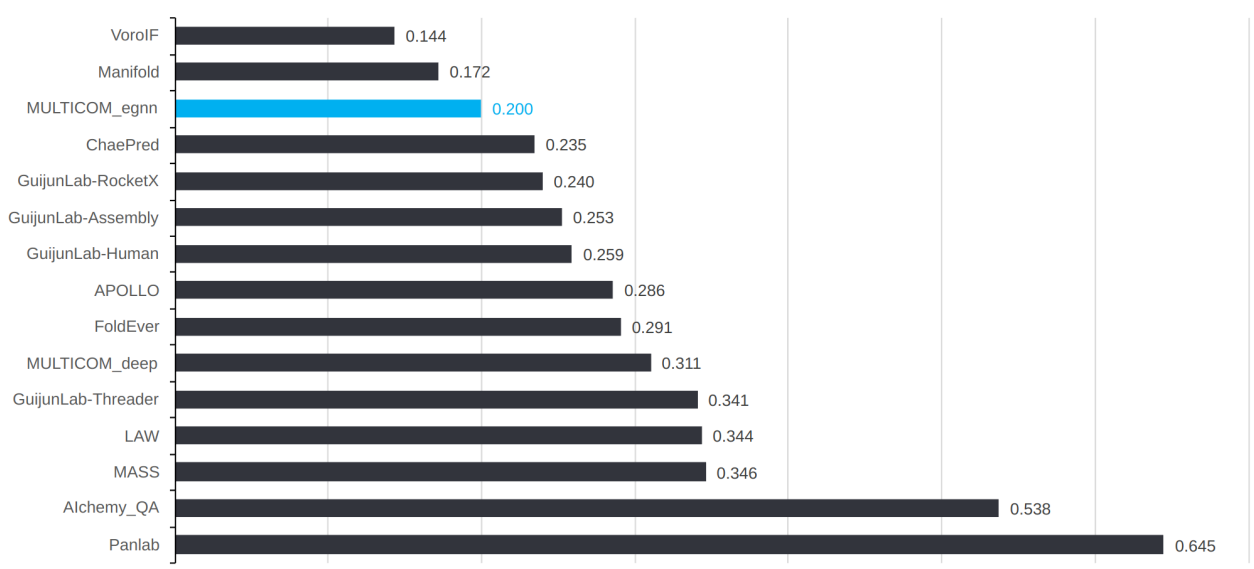

Fig. 3. The average TM-score ranking loss for all single-model methods. MULTICOM_egnn ranked 3rd among all single-model methods

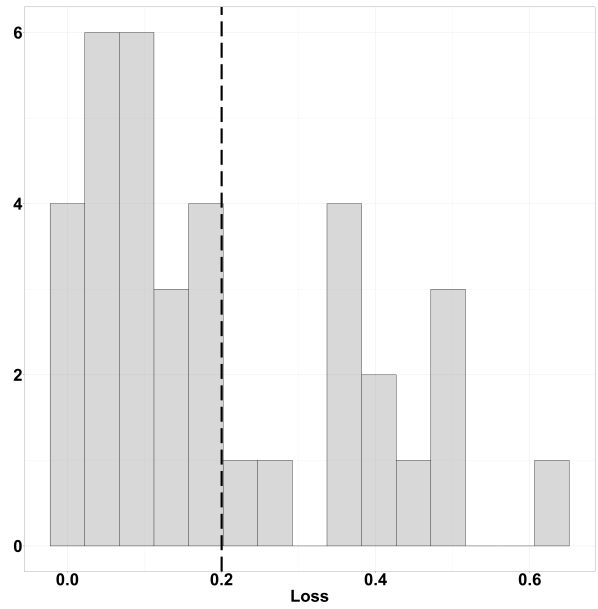

Fig. 4. MULTICOM_egnn's loss histogram for CASP15 targets. The black dashed vertical line represents the position of the mean value.

5 Discussion

Compared to the general high accuracy of predicted tertiary structures (Jumper *et al.* (2021)), the average accuracy of quaternary structures predicted for protein complexes (multimers) is still relatively low (Bryant *et al.* (2022)), making the selection of good models from a pool of decoys harder. We observe this not only in some popular complex datasets Lensink and Wodak (2014); Kundrotas *et al.* (2018), but also in our newly-built DBM55-AF2 sets. For instance, some targets like 6A0Z and 3U7Y in the DBM55-AF2 set only have a few decoys with acceptable or higher quality, while the rest of decoys have very low quality. If no model of acceptable or higher quality is ranked at the top, the ranking loss will be very high (see some examples in Table 1). Therefore, a significant challenge in ranking the models of protein multimers is to identify a few good models in a large pool of mostly bad models.

A more common way to evaluate the ranking ability of the quality assessment (QA) methods for protein complexes is the hit rate, a standard method used by CAPRI. However, a hit rate only measures the number of qualified decoys in the TOP N ranked models, without measuring the difference between the best possible model and the top-ranked model. Therefore, in this work, we also apply the loss metric widely used in evaluating the quality assessment methods for protein tertiary structures to the quality assessment for protein quaternary structures.

Considering these two metrics together helps us evaluate a protein multimeric QA method's ranking ability more effectively. For example, ZRANK2 which has slightly better hit-rate performance than DProQA on the DBM55-AF2 set, while its loss is much higher than DProQA's loss. Overall, DProQA demonstrates consistently good performance on our internal benchmark as well as the most rigorous blind CASP15 benchmark.

6 Conclusion

In this work, we present DProQA - a gated graph transformer for protein complex structure assessment. Our rigorous experiments and CASP15 results demonstrate that DProQA performs relatively well in ranking decoy models of protein complexes and the gated message passing in the transformer is useful for improving its performance. Both the tool and the new datasets consisting of multimer models predicted by AlphaFold2 and AlphaFold-Multimer are made publicly available for the community to further advance the field.

Funding

The project is partially supported by two NSF grants (DBI 1759934 and IIS 1763246), two NIH grants (R01GM093123 and R01GM146340), three DOE grants (DE-SC0020400, DE-AR0001213, and DE-SC0021303), and the computing allocation on the Summit compute cluster provided by Oak Ridge Leadership Computing Facility (Contract No. DE-AC05-00OR22725).

References

Athanasios, A. *et al.* (2017). Protein-protein interaction (ppi) network: recent advances in drug discovery. *Current drug metabolism*, **18**(1), 5–10.

Baker, D. (2006). Prediction and design of macromolecular structures and interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **361**(1467), 459–463.

Basu, S. and Wallner, B. (2016a). Dockq: a quality measure for protein-protein docking models. *PloS one*, **11**(8), e0161879.

Basu, S. and Wallner, B. (2016b). Finding correct protein–protein docking models using proqdock. *Bioinformatics*, **32**(12), i262–i270.

Bryant, P. *et al.* (2022). Improved prediction of protein-protein interactions using alphafold2. *Nature Communications*, **13**(1), 1–11.

Cao, Y. and Shen, Y. (2020). Energy-based graph convolutional networks for scoring protein docking models. *Proteins: Structure, Function, and Bioinformatics*, **88**(8), 1091–1099.

CASP15 (2022). 15th community wide experiment on the critical assessment of techniques for protein structure prediction. *CASP15*.

Chen, C. *et al.* (2023). 3d-equivariant graph neural networks for protein model quality assessment. *Bioinformatics*.

Chen, X. and Cheng, J. (2022). Distema: distance map-based estimation of single protein model accuracy with attentive 2d convolutional neural network. *BMC bioinformatics*, **23**(3), 1–14.

Chen, X. *et al.* (2020). Deep ranking in template-free protein structure prediction. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–10.

Chen, X. *et al.* (2021). Protein model accuracy estimation empowered by deep learning and inter-residue distance prediction in casp14. *Scientific Reports*, **11**(1), 1–12.

Dominguez, C. *et al.* (2003). Haddock: a protein- protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, **125**(7), 1731–1737.

Dwivedi, V. P. and Bresson, X. (2020). A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*.

Eismann, S. *et al.* (2021). Hierarchical, rotation-equivariant neural networks to select structural models of protein complexes. *Proteins: Structure, Function, and Bioinformatics*, **89**(5), 493–501.

Evans, R. *et al.* (2021). Protein complex prediction with alphafold-multimer. *BioRxiv*.

Gao, M. *et al.* (2021). High-performance deep learning toolbox for genome-scale prediction of protein structure and function. In *2021 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC)*, pages 46–57. IEEE.

Gao, M. *et al.* (2022). Proteome-scale deployment of protein structure prediction workflows on the summit supercomputer. *arXiv preprint*

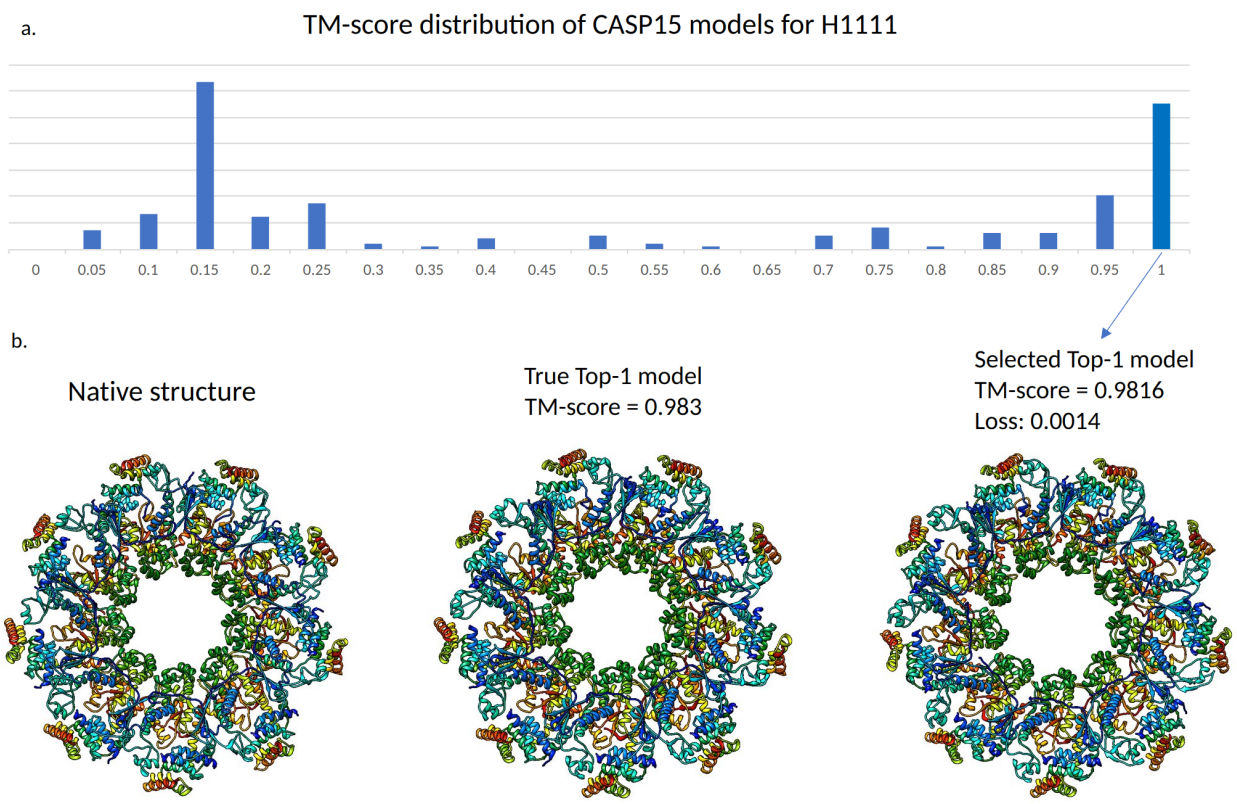

Fig. 5. Target H1111. (A) TM-score distribution of CASP15 models of H1111. (B). From left to right, the three protein complex structures shown are the corresponding native structure, the true TOP-1 model, and the MULTICOM_egnn top selected model, respectively. Here, MULTICOM_egnn achieved a 0.0014 TM-score ranking loss.

arXiv:2201.10024.

Geng, C. *et al.* (2020). iscore: a novel graph kernel-based function for scoring protein–protein docking models. *Bioinformatics*, **36**(1), 112–121.

Gray, J. J. *et al.* (2003). Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of molecular biology*, **331**(1), 281–299.

Guo, Z. *et al.* (2022). Prediction of inter-chain distance maps of protein complexes with 2d attention-based deep neural networks. *Nature Communications*, **13**(1), 1–10.

Han, Y. *et al.* (2021). Quality assessment of protein docking models based on graph neural network. *Frontiers in Bioinformatics*, **1**, 693211.

Hinton, G. E. *et al.* (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hopf, T. A. *et al.* (2019). The evcouplings python framework for coevolutionary sequence analysis. *Bioinformatics*, **35**(9), 1582–1584.

Hu, H. *et al.* (2019). Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3464–3473.

Huang, S.-Y. and Zou, X. (2008). An iterative knowledge-based scoring function for protein–protein recognition. *Proteins: Structure, Function, and Bioinformatics*, **72**(2), 557–579.

Ingraham, J. *et al.* (2019). Generative models for graph-based protein design. *Advances in neural information processing systems*, **32**.

Jing, B. *et al.* (2020). Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*.

Jumper, J. *et al.* (2021). Highly accurate protein structure prediction with alphafold. *Nature*, **596**(7873), 583–589.

Kinch, L. N. *et al.* (2021). Topology evaluation of models for difficult targets in the 14th round of the critical assessment of protein structure prediction (casp14). *Proteins: Structure, Function, and Bioinformatics*, **89**(12), 1673–1686.

Kortemme, T. and Baker, D. (2004). Computational design of protein–protein interactions. *Current opinion in chemical biology*, **8**(1), 91–97.

Kotthoff, I. *et al.* (2021). Dockground scoring benchmarks for protein docking. *Proteins: Structure, Function, and Bioinformatics*.

Kundrotas, P. J. *et al.* (2018). Dockground: a comprehensive data resource for modeling of protein complexes. *Protein Science*, **27**(1), 172–181.

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.

Lippow, S. M. and Tidor, B. (2007). Progress in computational protein design. *Current opinion in biotechnology*, **18**(4), 305–311.

Liu, S. *et al.* (2008). Dockground protein–protein docking decoy set. *Bioinformatics*, **24**(22), 2634–2635.

Liu, Z. *et al.* (2021). Swin transformer v2: Scaling up capacity and resolution. *arXiv preprint arXiv:2111.09883*.

Liu, Z. *et al.* (2022). Swin transformer v2: Scaling up capacity and resolution. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lundström, J. *et al.* (2001). Pcons: A neural-network–based consensus predictor that improves fold recognition. *Protein science*, **10**(11), 2354–2362.

Macalino, S. J. Y. *et al.* (2018). Evolution of in silico strategies for protein-protein interaction drug discovery. *Molecules*, **23**(8), 1963.

McGuffin, L. J. (2007). Benchmarking consensus model quality assessment for protein fold recognition. *BMC bioinformatics*, **8**(1), 1–15.

Mirdita, M. *et al.* (2021). Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics*, **37**(18), 3029–3031.

Moal, I. H. *et al.* (2013). The scoring of poses in protein-protein docking: current capabilities and future directions. *BMC bioinformatics*, **14**(1), 1–15.

Morehead, A. *et al.* (2021a). Dips-plus: The enhanced database of interacting protein structures for interface prediction. *arXiv preprint arXiv:2106.04362*.

Morehead, A. *et al.* (2021b). Geometric transformers for protein interface contact prediction. *arXiv preprint arXiv:2110.02423*.

Pierce, B. and Weng, Z. (2008). A combination of rescoring and refinement significantly improves protein docking performance. *Proteins: Structure, Function, and Bioinformatics*, **72**(1), 270–279.

Pierce, B. G. *et al.* (2011). Accelerating protein docking in zdock using an advanced 3d convolution library. *PloS one*, **6**(9), e24657.

Pons, C. *et al.* (2011). Scoring by intermolecular pairwise propensities of exposed residues (sipper): a new efficient potential for protein- protein docking. *Journal of chemical information and modeling*, **51**(2), 370–377.

Rao, R. M. *et al.* (2021). Msa transformer. In *International Conference on Machine Learning*, pages 8844–8856. PMLR.

Scott, D. E. *et al.* (2016). Small molecules, big targets: drug discovery faces the protein–protein interaction challenge. *Nature Reviews Drug Discovery*, **15**(8), 533–550.

Tovchigrechko, A. and Vakser, I. A. (2006). Gramm-x public web server for protein–protein docking. *Nucleic acids research*, **34**(suppl_2), W310–W314.

Uziela, K. and Wallner, B. (2016). Proq2: estimation of model accuracy implemented in rosetta. *Bioinformatics*, **32**(9), 1411–1413.

Uziela, K. *et al.* (2016). Proq3: Improved model quality assessments using rosetta energy terms. *Scientific reports*, **6**(1), 1–10.

Vaswani, A. *et al.* (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

Veličković, P. *et al.* (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Vreven, T. *et al.* (2011). Integrating atom-based and residue-based scoring functions for protein–protein docking. *Protein Science*, **20**(9), 1576–1586.

Vreven, T. *et al.* (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.

Wang, X. *et al.* (2020). Protein docking model evaluation by 3d deep convolutional neural networks. *Bioinformatics*, **36**(7), 2113–2118.

Wang, X. *et al.* (2021). Protein docking model evaluation by graph neural networks. *Frontiers in Molecular Biosciences*, **8**, 402.

Wu, T. *et al.* (2021). Deepdist: real-value inter-residue distance prediction with deep residual convolutional network. *BMC bioinformatics*, **22**(1), 1–17.

Xu, B. *et al.* (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

Yan, Y. and Huang, S.-Y. (2021). Accurate prediction of inter-protein residue–residue contacts for homo-oligomeric protein complexes. *Briefings in bioinformatics*, **22**(5), bbab038.

Zhang, C. *et al.* (2022). Us-align: Universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *bioRxiv*.

Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, **57**(4), 702–710.

Zhou, H. and Skolnick, J. (2011). Goap: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophysical journal*, **101**(8), 2043–2052.

Zhou, H. and Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein science*, **11**(11), 2714–2726.