Computational and Structural Biotechnology Journal 27 (2025) 160–170

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

# EquiRank: Improved protein-protein interface quality estimation using protein language-model-informed equivariant graph neural networks

Md Hossain Shuvo [a], Debswapna Bhattacharya [b,*]

[a] Department of Computer Science, Prairie View A&M University, Prairie View, 77446, TX, USA
[b] Department of Computer Science, Virginia Tech, Blacksburg, 24061, VA, USA

ARTICLE INFO

Dataset link: https://dx.doi.org/10.5281/zenodo.7841307, https://predictioncenter.org/download_area/CASP13/predictions/oligo/, https://predictioncenter.org/download_area/CASP14/predictions/oligo/, https://predictioncenter.org/download_area/CASP15/predictions/oligo/, https://dockground.compbio.ku.edu/downloads/unbound/decoy/decoys1.0.zip

*Keywords:*
Protein-protein interaction
Protein complex quality estimation
Protein language models
Graph neural networks
Deep learning

ABSTRACT

Quality estimation of the predicted interaction interface of protein complex structural models is not only important for complex model evaluation and selection but also useful for protein-protein docking. Despite recent progress fueled by symmetry-aware deep learning architectures and pretrained protein language models (pLMs), existing methods for estimating protein complex quality have yet to fully exploit the collective potentials of these advances for accurate estimation of protein-protein interface. Here we present EquiRank, an improved protein-protein interface quality estimation method by leveraging the strength of a symmetry-aware E(3) equivariant deep graph neural network (EGNN) and integrating pLM embeddings from the pretrained ESM-2 model. Our method estimates the quality of the protein-protein interface through an effective graph-based representation of interacting residue pairs, incorporating a diverse set of features, including ESM-2 embeddings, and then by learning the representation using symmetry-aware EGNNs. Our experimental results demonstrate improved ranking performance on diverse datasets over existing latest protein complex quality estimation methods including the top-performing CASP15 protein complex quality estimation method VoroIF_GNN and the self-assessment module of AlphaFold-Multimer repurposed for protein complex scoring and across different performance evaluation metrics. Additionally, our ablation studies demonstrate the contributions of both pLMs and the equivariant nature of EGNN for improved protein-protein interface quality estimation performance. EquiRank is freely available at https://github.com/mhshuvo1/EquiRank.

## 1. Introduction

Protein-protein interactions play a fundamental role in driving a wide range of biological processes [1–3]. Although there has been a significant leap toward accurately predicting the monomeric protein models by AlphaFold 2 [4], the accurate prediction of their interactions still remains challenging [5–8]. Existing protein complex modeling methods typically generate alternative complex models (a.k.a. decoys) by performing conformational sampling of the protein-protein interaction interfaces [9–12]. As such, the scoring of protein-protein interfaces for the identification of the most accurate conformation is, therefore, an important component of a successful protein complex structure modeling process [13,14].

There has been promising progress in the development of various methods for estimating the accuracy scores of protein complexes with the application of various machine learning models. For instance, TRScore [15] uses the ResNet-inspired [16] VGG network to learn the voxelized representation of protein complexes for scoring protein complexes. Similarly, DOVE [17] employs Convolutional Neural Network (CNN) [18] to score protein complexes, utilizing 3D voxelized representation integrated with atomic energy. PIsTon [19] uses a vision transformer (ViT) with empirical-based energy terms for evaluating Protein binding Interfaces. Recent advances in Graph Neural Network (GNN)-based models have gained attention for their ability to effectively represent molecular structures [20,21] thus prompting their application to protein complex quality estimation problems. Methods that utilize Graph Neural Networks typically represent a protein-protein interface as a graph, with each residue as a node and their interactions as edges in the graph. For instance, our recent method PIQLE [22] successfully applies the Graph Attention Network [23] to estimate the accuracy of the protein-protein interface. VoroIF_GNN [24] estimates the precision of protein complexes using the attention-based graph neural network to learn the atom-level graph derived from the Voronoi tessellation-based interface representations. EuDockScore [25] uses an euclidean graph

* Corresponding author.
*E-mail address:* dbhattacharya@vt.edu (D. Bhattacharya).

https://doi.org/10.1016/j.csbj.2024.12.015
Received 31 October 2024; Received in revised form 18 December 2024; Accepted 20 December 2024

Available online 30 December 2024

neural network score to assess the protein-protein interface. DProQA [26] employs a graph transformer network for estimating the quality of protein complexes. GDockScore [27] utilizes a bi-directional graph attention network for estimating the quality score of protein complexes. GNN-DOVE [17] represents the protein interface as a graph and employs Graph Attention Network (GAT) to estimate the protein complex score.

In addition to exploiting the graph representations of protein complexes with several representative features, latest protein complex quality estimation methods leverage the power of protein language models (PLMs), which have revolutionized diverse predictive tasks, including protein structure prediction and protein function prediction in recent years [28–36] For instance, DeepRank-GNN-esm [37] uses a graph neural network (GNN) that incorporates embeddings from the transformer-based protein language model ESM-2 [31] to better capture the graph representation of a protein complex. Despite the success of existing protein complex quality estimation methods in applying GNN-based approaches, such as Graph Attention Network (GAT) [23] and Graph Transformer Network (GTN) [38], that operate efficiently on protein graphs lacking a fixed order and providing important invariant properties regardless of node permutation, they do not inherently capture symmetry under certain rotations and translations while dealing with such 3D objects such as proteins [39,40]. Therefore, considering the 3D structure of proteins, Equivariant Neural Network (EGNN) is desirable that incorporates the spatial location of each residue [40,41], providing properties such as equivariance to rotation and translation, while also maintaining invariance to node permutation. Furthermore, the combination of embeddings from the Protein Language Model for efficiently representing nodes in the interaction graph and EGNN for learning the representation enhances the overall robustness of the framework [42].

Here we present, an improved protein-protein interface quality estimation method EquiRank. EquiRank introduces several advances over existing protein complex quality estimation methods and our previous work PIQLE [22] including: i) the application of Symmetry-aware Equivariant Graph Neural Network (EGNN) [40], ii) training of an Ensemble of four EGNN to separately learn multimeric distance and orientation representations, iii) integration of the embeddings from the protein language model based ESM-2 model, and iv) comprehensive benchmarking and validation on diverse set of datasets.

Starting from a given protein complex, EquiRank extracts the interface graph of interacting residue pairs and combines various representative features including the embeddings from the protein language model based ESM-2 model and our novel multimeric geometries, as employed in our recently published protein-protein interface quality estimation method PIQLE [22]. EquiRank delivers improved protein-protein interface quality estimation performance over state-of-the-art protein complex quality estimation methods across diverse datasets and various accuracy measures. Our ablation study reveals that the improved performance of EquiRank is directly connected to the integration of protein language model-based embeddings and EGNN. EquiRank is freely available at https://github.com/mhshuvo1/EquiRank.

## 2. Material and methods

### 2.1. Features generation

*Interface graph representation:* We represent a protein complex as an interface graph G = (V, E) consisting of interacting residue pairs with each of the residues in the interface as a node v ∈ V and an interacting residue pair as an edge e ∈ E as shown in Fig. 1A. We extract the interface residue pairs from a given protein complex based on the inter-residue distances of their C$\beta$ (C$\alpha$ for glycine) atoms. Specifically, we define a residue pair as interactive when the distance between their C$\beta$ (C$\alpha$ for glycine) atoms is less than 10Å [22]. We represent each of the nodes in the interface graph with 349 features including protein language model (pLM)-based and AlphaFold 2 distilled Multiple Sequence Alignment (MSA) features. Additionally, we represent each of the edges with multimeric distance and orientation having 29 features.

#### 2.1.1. Node feature:

We generate a total of 349 features including both sequence- and structure-based features for representing each of the nodes in the interface graph.

1. Protein Language Model (pLM) based features (33): We use a variant of the pre-trained ESM-2 model [31] with 650M parameters and 33 layers, producing sequence embeddings with a shape of L × 33, where L is the length of the sequence. This model generates embeddings for each amino acid sequence of the interacting monomers in the protein complex. Afterward, we perform a sigmoidal transformation on the resulting embeddings to generate 33 Protein Language Model (pLM)-based features.
2. Multiple sequence alignment features (256): We employ ColabFold [43] to generate Multiple Sequence Alignment (MSA) using MMseq2 [44] for each of the amino acid sequences in the protein complex. The generated MSA is then input to the EvoFormer blocks of AlphaFold 2 [4], implemented in ColabFold, producing distilled MSA representations encoded as a dictionary. Subsequently, we extract the first row of the distilled MSA representation ("msa_first_row" from the dictionary) and then apply a sigmoidal transformation to generate 256 MSA features.
3. Evolutionarily features (2): We generate evolutionarily features in the form of the number of effective sequences (Neff) computed from the Multiple Sequence Alignment (MSA) of the individual amino acid sequences within the protein complex and its concatenated MSA, by following a similar methodology as adopted in PIQLE [45]. The number of effective sequences (Neff) represents the depth of the MSA, thereby considering the evolutionarily information.
4. Amino acid encoding (21): We represent each of the nodes in the interface graph, corresponding to a specific amino acid residue, using a one-hot encoded binary vector, consisting of 20 naturally occurring amino acid types and a gap for non-standard amino acid [46].
5. Relative residue positioning (1): We obtain the relative positional information for each node in the interface graph corresponding to each of the residues in the sequence of a model as follows,

$$\mathrm{relPos}(aa) = \frac{\mathrm{aa}^n}{L} \tag{1}$$

Where $\mathrm{aa}^n$ is the $n$-th amino acid residue in the sequence and $L$ is the length of the sequence.
6. Secondary structure and solvent accessibility (13): We use the DSSP [47] program to generate secondary structure and solvent accessibility from the structure. For each of the residues corresponding to the nodes in the graph, we generate a binary vector of one-hot encoding for 8-state secondary structure types, resulting in 8 secondary structure features. Additionally, we transform the 8-state secondary structure into a 3-state by grouping them into helices, strands, and coils [22]. Subsequently, we generate a one-hot encoded binary vector of 3-state secondary structures, resulting in 3 features. We discretize real-valued solvent accessibility into buried and exposed [45] and generate 2 features by performing one-hot encoding of the corresponding types.
7. Local backbone geometry (4): To capture the local backbone geometry, we calculate phi ($\phi$) and psi ($\psi$) backbone torsion angles from the structure to capture the local backbone geometry of each residue. Subsequently, we generate 4 features by performing sinusoidal and cosine transformations of the angles [48].
8. Ultra shape recognition features (3): To capture the topological relationship between the residue and the overall structure, we calculate residue-level Ultra Shape Recognition (USR) [49] features. Following a similar approach as adopted in DeepUMQA [50], we compute the residue-level USR feature representing the spatial relationship

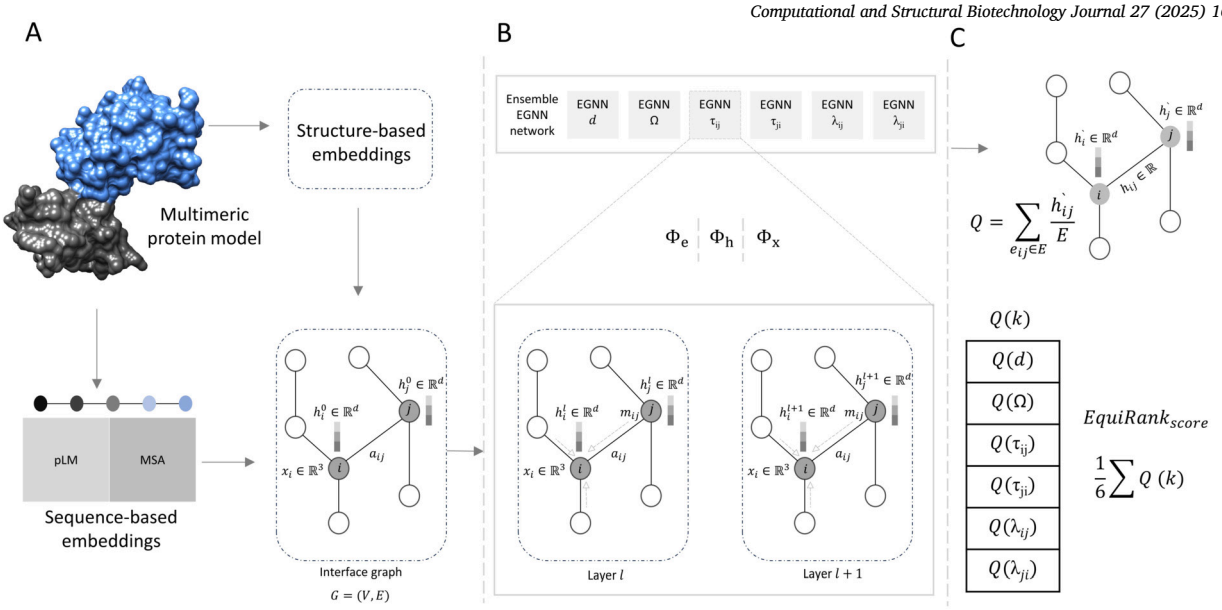


**Fig. 1.** Flowchart of EquiRank framework for protein-protein interface quality estimation. A) Generation of sequence- and structure-based features, including pLM-based sequence embeddings and MSA-based encoding, and multimeric distance and orientation, respectively, from the predicted protein complex structure with two interacting monomers colored in blue and gray. B) Architecture of ensemble Equivariant Graph Neural Network (EGNN). C) Edge-level regression of interacting residue pairs for transformed distance ($d$) and normalized angular RMSD of multimeric orientations Ω, $\lambda_{12}$, $\lambda_{21}$, $\tau_{21}$, $\tau_{21}$, and their probabilistic combination for estimating the protein-protein interface quality.

information between the structure and each specific residue using three residue distance sets. Subsequently, we apply min-max transformation on all three distance sets to generate normalized USR features for each residue of a corresponding node in the interface graph, resulting in 3 USR features.

9. Residue orientation (3): To define the orientation of each amino acid residue, we derive three features for each residue, corresponding to a node in the graph. These features include two features, calculated as the unit vectors in the directions of $C_{\alpha_{i+1}} - C_{\alpha_i}$ and $C_{\alpha_{i-1}} - C_{\alpha_i}$, and one feature, calculated as the unit vector of $C_{\beta_i} - C_{\alpha_i}$, based on the assumption of tetrahedral geometry [51].
10. Residue neighbors (1): We calculate the number of spatial neighbors of a node in the interface graph in terms of the corresponding monomeric structure where two residues are considered to be neighbors if the intra-residue distance of their C$\beta$ (C$\alpha$ for glycine) atoms is less than 10Å. Once again, we perform a min-max transformation to generate the normalized number of neighbors, resulting in 1 feature.
11. Rosetta centroid energy terms (12): We generate 12 Rosetta [52, 53] centroid energy terms following a similar approach as used in QDeep [54]. For each of the nodes in the interface graph, we use the sigmoidal transformation of the energy terms for the corresponding residue in the structure and use them as 12 energy-based features.

*2.1.2. Edge features:*

For each of the edges in the interface graph, we generate a total of 29 edge features as outlined below:

1. Multimeric interaction distance (17): For each of the edges in the interface graph, we calculate the Euclidean distance between the $C_\beta$ ($C_\alpha$ for glycine) atoms of the corresponding interacting residue pairs. We first discretized the calculated distance from $\leq 0.2$ to $< 10$ Å into 17 bins, each having a uniform bin width of 0.5Å. Afterward, we use the one-hot encoding of the discretized bin, resulting in 17 multimeric interaction distance-based edge features.
2. Multimeric orientation (10): To capture the orientation information between the interacting residue pairs, we extend the work of trRosetta [55] for multimers to represent the edges in the interface graph with orientation features. Specifically, each edge is represented by 3 torsion angles (Ω, $\tau_{ij}$, $\tau_{ji}$) and 2 planar angles ($\lambda_{ij}$, $\lambda_{ji}$) [22]. The Ω torsion angle measures rotation along the virtual axis connecting the $C_\beta$ atoms of the interacting interface residue pairs, and $\tau_{ij}$, $\lambda_{ij}$ ($\tau_{ji}$, $\lambda_{ji}$) angles specify the direction of the $C_\beta$ atom of the interface residue from the first (second) interacting monomer in a reference frame centered on the interface residue from the second (first) interacting monomer. Unlike the symmetric torsion angle Ω, $\tau$ and $\lambda$ are asymmetric and depend on the order of the interacting residue pairs. Subsequently, we perform sinusoidal and cosine transformations of the angles, leading to 10 features.
3. Relative positioning of the interacting residues (2): To capture the positional information of interacting residue pairs in the protein complex, we obtain the relative positional information for each of the interacting residue pairs as follows,

$$\text{relPos(aa)} = \left| \frac{\text{aa}_i^n}{L} - \frac{\text{aa}_j^n}{L} \right| \quad (2)$$

Where $\text{aa}_i^n$ and $\text{aa}_j^n$ are the $n$th interacting residues in the structure, and $L$ is the length of the sequence.

4. Neighbors of interacting residue pairs (1): We calculate the number of spatial neighbors of each of the interacting residue pairs in terms of the corresponding monomeric structure where two residues are considered to be neighbors if the intra-residue distance of their C$\beta$ (C$\alpha$ for glycine) atoms is less than 10Å. As aforementioned, we perform min-max transformation to generate the normalized number of neighbors for each interacting residue and subsequently use their weighted combination, resulting in 1 feature.

*2.2. Dataset*

Table 1 shows the datasets for training, testing, and validating the ensemble EGNN models in EquiRank. To train the models, we collect a total of 18,022 protein complex models for a non-redundant set of 1,127 dimeric targets having lengths ranging from 67 to 1,375. We first collect 14,400 models for 1,097 heterodimeric targets from VoroIF_GNN [24] (hereafter called VoroIF_GNN_train). We additionally incorporate 3,622 models for 30 dimer targets from both CASP13 and CASP14 in the

**Table 1**
Distribution of training, testing, and validation datasets.

| | Dataset | Num targets | Num decoys | Correct (DockQ >0.23) | Incorrect (DockQ <0.23) |
|---|---|---|---|---|---|
| Training | VoroIF_GNN_train | 1,097 | 14,400 | 42% | 58% |
| Training | CASP13 | 20 | 2384 | 17.24% | 82.76% |
| Training | CASP14 | 10 | 1238 | 15.95% | 84.05% |
| Testing | VoroIF_GNN_test | 235 | 2845 | 42% | 58% |
| Testing | Dockground v1 | 23 | 2500 | 10.72% | 89.28% |
| Testing | CASP15 | 26 | 6850 | 45.37% | 54.63% |
| Validation | VoroIF_GNN_validation | 235 | 2,814 | 40.96% | 59.14% |

**Table 2**
Pairwise sequence identity between training, testing, and validation datasets.

| Datasets | VoroIF_GNN_train[a] | VoroIF_GNN_test[b] | CASP13[a] | CASP14[a] | CASP15[b] | Dockgroundv1[b] | VoroIF_GNN_validation[c] |
|---|---|---|---|---|---|---|---|
| Voroif_GNN_train | | 19.57% | 21.28% | 22.28% | 20.08% | 19.83% | 19.29% |
| VoroIF_GNN_test | | | 21.41% | 22.37% | 20.16% | 19.98% | 19.40% |
| CASP13 | | | | 20.0% | 17.64% | 17.53% | 21.35% |
| CASP14 | | | | | 15.27% | 15.09% | 22.42% |
| CASP15 | | | | | | 19.53% | 20.08% |
| Dockgroundv1 | | | | | | | 19.97% |
| VoroIf_GNN_validation | | | | | | | |

[a] Training datasets
[b] Testing datasets
[c] Validation datasets.

training set. Subsequently, we train our model by combining the three datasets, VoroIF_GNN_train, CASP13, and CASP14. We benchmark the performance of our method EquiRank and other competing methods on a set of 12,195 models for 284 targets having lengths ranging from 29 to 1,677 and diverse accuracy with both correct and incorrect protein complex models. We use a DockQ threshold of 0.23 to differentiate between correct and incorrect models. First, we collect 2,845 models for 235 targets from VoroIF_GNN (hereafter called VoroIF_GNN_test) having 42% correct and 58% incorrect protein complex models. Additionally, we collected 2,500 models for 23 targets from Dockground version 1 [56] (hereafter called Dockground v1) consisting of 10.72% correct models and 89.28% incorrect complex models. Finally, we collect CASP15 datasets for protein complex targets having 6,850 decoys for 26 targets with 45.37% correct and 54.63% incorrect decoys. Additionally, for ablation studies, we collect 2,814 models for 235 targets from VoroIF_GNN (hereafter called VoroIF_GNN_validation) having a model length ranging from 96 to 1,222. It is noteworthy that all the datasets used for training, benchmarking, and ablation studies are non-overlapping with an average pairwise sequence identity of <23% between any pair of datasets as shown in Table 2.

### 2.3. Evaluation metrics and competing methods

We evaluate the performance of our method EquiRank in terms of various evaluation metrics. For all performance measures, we use the DockQ [57] score as ground truth. DockQ score quantifies the quality of protein-protein docking models by incorporating measures such as $F_{Nat}$, LRMS, and iRMS [58]. $F_{Nat}$ is the proportion of native interfacial contacts preserved in the docking model. LRMS and iRMS are root mean square deviations for the ligand and interface, respectively, calculated with specific scaling factors optimized to distinguish docking models according to CAPRI standards. These scaling factors are set at 8.5 Å for LRMS and 1.5 Å for iRMS to maximize F1 scores in classification. The DockQ score is a weighted combination of multiple scoring terms and is calculated as follows:

$$\mathrm{RMS}_{\mathrm{scaled}}(\mathrm{RMS}, d_i) = \frac{1}{1 + \left(\frac{\mathrm{RMS}}{d_i}\right)^2}, \tag{3}$$

$$\mathrm{DockQ}(F_{\mathrm{nat}}, \mathrm{LRMS}, \mathrm{iRMS}, d_1, d_2) = \frac{F_{\mathrm{nat}} + \mathrm{RMS}_{\mathrm{scaled}}(\mathrm{LRMS}, d_1) + \mathrm{RMS}_{\mathrm{scaled}}(\mathrm{iRMS}, d_2)}{3}. \tag{4}$$

To evaluate the methods' ranking ability, we use per-target Spearman correlation coefficients between the decoys' predicted and the true DockQ scores, calculated as follows:

$$\overline{\rho}_{\mathrm{Spearman}} = \frac{1}{N}\sum_{i=1}^{N} \rho_{\mathrm{Spearman}}(Predicted, DockQ), \tag{5}$$

Where $Predicted$ and $DockQ$ refer to the decoys' predicted and corresponding true scores for a specific target $i$, respectively and N refers to the number of targets. A higher correlation indicates a better ranking ability of a method. Additionally, we evaluate methods' ranking performance in terms of top-N hit rate and success rate [15]. Top-N hit rate is the fraction of acceptable models among top-ranked models relative to all acceptable models in a specific dataset and is calculated as follows:

$$\text{Hit rate}(N) = \frac{H(N)}{M} \times 100\% \tag{6}$$

where $H(N)$ represents the number of acceptable models with a DockQ threshold of 0.23, among top-N ranked models and $M$ represents the total number of acceptable models in the corresponding dataset. A higher top-N hit rate indicates better ranking ability. On the other hand, success rate is the percentage of targets with at least one acceptable model among top-N ranked models and is calculated as follows:

$$\text{Success rate}(N) = \frac{S(N)}{K} \times 100\% \tag{7}$$

where $S(N)$ represents the number of targets having at least one acceptable models with a DockQ threshold of 0.23 among top-N ranked models and $K$ represents the total number of targets. Similarly, a higher success rate indicates better ranking ability. For the VoroIF_GNN_test dataset, we calculate the Top-1, Top-5, Top-10, and Top-15 hit rates and success rates due to the limited number of models per target. For the Dockground v1 dataset, we compute the Top-1, Top-5, Top-10, Top-15, Top-20, Top-25, and Top-30 hit rates and success rates.

Additionally, to evaluate the methods' ability to accurately distinguish between high-quality models and others, we report the area under the ROC curve (AUC) with a DockQ threshold of 0.80 [57]. The Area Under the Curve (AUC) is calculated by plotting the true positive rate (TPR) against the false positive rate (FPR) [59] for a DockQ threshold of 0.8. A higher AUC value indicates better distinguishability of a method in separating high-quality models. We also calculate the Precision-Recall Area Under the Curve (PRAUC) with the same DockQ threshold of 0.80. PRAUC is calculated by plotting Precision on the y-axis and Recall on

the x-axis [59] for DockQ thresholds of 0.8. A higher PRAUC indicates improved distinguishability in separating high-quality models.

We compare the performance of EquiRank against state-of-the-art existing complex quality prediction methods using the same evaluation metric. We use several competing methods, leveraging diverse deep learning architectures. We compare EquiRank against graph neural network-based methods VoroIF_GNN [24] PIQLE [22], DProQA [26], GNN-DOVE [17], GDockScore [27], DeepRank-GNN-esm [37] and EuDockScore [25]. We evaluate EuDockScore using its provided predicted scores for the CASP15 dataset in terms of its ability to distinguish high-quality models. Additionally, we compare our method with Convolutional Neural Network (CNN)-based methods such as variants of DOVE [60] and TRScore [15]. Moreover, we compare our method with the interface-predicted TM scores (iPTM) predicted by the self-assessment module of AlpaFold-Multimer [7]. Specifically, we utilize an extended version of the AF2Rank [61] method repurposed for estimating the protein complex quality based on the self-assessment module of AlphaFold-multimer.

### 2.4. Network architecture

We employ a deeper Equivariant Graph Neural Network (EGNN) [40] to estimate the quality of the protein-protein interface as shown in Fig. 1B. Our deep EGNN consists of four stacked convolutional layers (EGNNConv), operating on the protein-protein interface graph. It accepts both node ($h_n$) and edge ($h_e$) features, as well as the Cartesian coordinate information ($h_x$) of the C$\beta$ (Ca for glycine) atoms of the interacting residue pairs $(i, j)$. Each EGNNConv performs a series of operations on the edge, node, and coordinate features of the interface graph, denoted by $\Phi_e$, $\Phi_h$, $\Phi_x$ respectively. Each operation consists of two-layer Multi-Layer Perceptrons (MLPs).

EGNNConv starts processing by performing edge embeddings for interacting residue pairs using a message-passing operation on the edges ($m_{ij}$) of the interacting residue pairs. This is done by applying a two-layer MLP, $\Phi_e$, on the node features ($h_n$) and edge features ($a_{ij}$), while considering the coordinates $(x_i, x_j)$ features as follows,

$$m_{ij} = \Phi_e(h_i^l, h_j^l, \|x_i^l - x_j^l\|^2, a_{ij}) \tag{8}$$

Where $m_{ij}$ is the edge message between the interacting residue pairs $i$ and $j$, and $h_i^l$ and $h_j^l$ are their node features at layer $l$. $\|x_i^l - x_j^l\|^2$ is the Euclidean distance between the Cartesian coordinates of the interacting residue pairs $i$ and $j$, a fundamental operation for implementing equivariant message passing operation within EGNN, setting it apart from the traditional Graph Neural Network (GNN).

Afterward, the edge embedding $m_{ij}$ from the previous layer is used to update the coordinates of residue $i$ ($x_i$) in the interacting residue pairs in the next layer $(l + 1)$. Specifically, the coordinates of $i$ ($x_i$) are updated by the sum of all relative differences of interacting residue pairs weighted by the edge embedding $m_{ij}$ as follows:

$$x_i^{(l+1)} = x_i^l + C \sum_{j \neq i} \left(x_i^l - x_j^l\right) \Phi_x(m_{ij}) \tag{9}$$

Where $x_i^l$ is the coordinates of residue $i$ in the interacting residue pairs and $(x_i^l - x_j^l)$ is the difference between their coordinates, $m_{ij}$ is the edge embedding from the previous layer. $C$ is a normalizing constant computed as

$$\left(\frac{1}{N(i)}\right) \tag{10}$$

where $N$ is the number of neighbors of residue $i$.

Afterward, the network aggregates all the messages from all the neighboring nodes of $i$ to update its features as follows,

$$m_i = \sum_{j \neq i} m_{ij} \tag{11}$$

Finally, a non-linear transformation is applied to the aggregated message and the node feature of $i$ in the current layer ($h_i^l$), producing the node embeddings with the updated node features of residue $i$ in the next layer $(l + 1)$ as follows,

$$h_i^{(l+1)} = \Phi_h(h_i^l, m_i) \tag{12}$$

### 2.5. Model training

To train our ensemble EGNN models, we assign the ground truth interface quality scores to each of the interacting residue pairs, representing edges $e_{ij}$ in the interface graph. Specifically, we generate 6 sets of features by assigning ground truth multimeric geometry, including one multimeric distance and five orientation labels. We calculate the multimeric distance label by employing a similar approach as adopted in PIQLE [22], where we first calculate the observed $C_\beta - C_\beta$ distance between the interacting interface residue pairs in the predicted complex structural model ($d_{ij}^{\text{model}}$) and the corresponding residue pairs in the native structure ($d_{ij}^{\text{native}}$). We then assign a normalized ground truth distance score $z_{ij}(d)$ to the edge $e_{ij}$ as follows:

$$z_{ij}(d) = \begin{cases} 1 & \text{if } d_{ij}^{\text{model}} < 10\,\text{Å and } d_{ij}^{\text{native}} < 10\,\text{Å} \\ \dfrac{1}{1+\left(\dfrac{\left|d_{ij}^{\text{model}} - d_{ij}^{\text{native}}\right|}{d_0}\right)^2} & \text{otherwise} \end{cases} \tag{13}$$

where $\left|d_{ij}^{\text{model}} - d_{ij}^{\text{native}}\right|$ is the observed edge-level error between the interacting interface residue pairs corresponding to the edge $e_{ij}$, and $d_0$ is a normalizing constant whose value is set to 10 Å.

Additionally, to learn the orientation error, we assign multimeric orientation for each edge $e_{ij}$ by calculating the normalized angular RMSD between torsion $(\Omega, \tau_{ij}, \tau_{ji})$ and planar angles $(\lambda_{ij}, \lambda_{ji})$ of the interacting residue pairs in the predicted complex structural model ($d_{ij}^{\text{model}}$), calculated by extending the work of trRosetta [55] for multimers and the corresponding residue pairs in the native structure ($d_{ij}^{\text{native}}$).

$$z_{ij}(\vec{a}) = \sqrt{\left(\min\left(\left|a_{ij}^{\text{native}} - a_{ij}^{\text{model}}\right|, 2\pi - \left|a_{ij}^{\text{native}} - a_{ij}^{\text{model}}\right|\right)\right)^2} \tag{14}$$

where $\vec{a}$ represents the torsion angles $\Omega$, $\tau_{ij}$, $\tau_{ji}$, planar angles $\lambda_{ij}$, $\lambda_{ji}$. Once again, $\tau$ and $\Omega$ are asymmetric and therefore depend on the order of the residues. Afterward, we normalize the angular RMSD for the torsion angles $\Omega$ and $\tau$ as follows:

$$\text{Normalized Angular RMSD} = \frac{1}{1+\left(\dfrac{z_{ij}(\vec{a})}{\frac{\pi}{4}}\right)^2} \tag{15}$$

where $\vec{a} \in \{\tau_{ij}, \tau_{ji}, \Omega\}$.

However, we use more stringent normalization criteria for the planar angle $z_{ij}(\lambda)$ as follows:

$$\text{Normalized Angular RMSD} = \frac{1}{1+\left(\dfrac{z_{ij}(\lambda)}{\frac{\pi}{8}}\right)^2} \tag{16}$$

Both $z_{ij}(d)$ and $z_{ij}(\Omega, \lambda_{ij}, \lambda_{ji}, \tau_{ij}, \tau_{ji})$ range between 0 to 1, with a higher score indicating better similarity between the interacting residue pairs in the model and the corresponding residue pairs in the native structure.

Therefore, we train our ensemble EGNN models on these 6 sets of features to independently learn $z_{ij}(d)$ and $z_{ij}(\Omega, \lambda_{ij}, \lambda_{ji}, \tau_{ij}, \tau_{ji})$ through edge-level error regression by optimizing the mean squared error loss function with sum reduction using the Deep Graph Library [62]. We use the Adam optimizer [63] with a learning rate of 0.001 and a weight decay of 0.0005. We train the model using an average batch size of 98, calculated by dividing the total number of edges across all graphs by the total number of graphs, to ensure consistent and manageable graph sizes

during training. The training process consists of at most 500 epochs on an NVIDIA A100 GPU, using an early stopping criterion based on the validation loss, where training stops if the loss does not decrease for 40 consecutive epochs (patience = 40) to prevent overfitting.

### 2.6. Estimation of protein-protein interface quality:

Fig. 1C shows the estimation of the protein-protein interface quality in two steps. Using each of the trained models, we first estimate the embeddings $h_i^l$ and $h_j^l$ for each of the interacting residue pairs $i$ and $j$ respectively. Afterward, we perform a dot product between the estimated embeddings $(h_i^l, h_j^l)$ for each of the interacting residue pairs to individually estimate their quality as follows,

$$h'_{ij} = h_i^l \cdot h_j^l \tag{17}$$

Where $h_i^l$ and $h_j^l$ represent the node embeddings of the interacting residue pairs $i$ and $j$ respectively, and $h'_{ij}$ is their estimated local quality score. For the global quality score, we first perform a probabilistic combination of the local quality of all the interacting residue pairs $|e|$ in the interface graph, estimated by each of the trained models as follows:

$$Q(k) = \frac{\sum_{e_{ij} \in E} h'_{ij}}{|E|} \quad \text{where } k \in \{z_{ij(d)}, \Omega, \lambda_{ij}, \lambda_{ji}, \tau_{ij}, \tau_{ji}\} \tag{18}$$

Where $|E|$ is the total number of edges in the interface graph, and $Q$ is the global quality, ranging between 0 and 1 from each of the ensemble models. Afterward, we estimate the EquiRank_score by performing an ensemble averaging of the global quality from each of the models trained with ground truth multimeric distance, $z_{ij}$ and geometry $\Omega, \lambda_{ij}, \lambda_{ji}, \tau_{ij}, \tau_{ji}$ as follows,

$$\text{EquiRank\_score} = \frac{Q(d) + Q(\Omega) + Q(\lambda_{ij}) + Q(\lambda_{ji}) + Q(\tau_{ij}) + Q(\tau_{ji})}{6} \tag{19}$$

Where EquiRank_score ranges between 0 and 1, with a higher score indicating better protein-protein interface quality.

## 3. Results and discussion

### 3.1. Ability to rank predicted models

Fig. 2 shows the ranking performance of EquiRank and other competing methods in terms of per-target average Spearman correlation between the methods' predicted and ground truth DockQ [57] scores on VoroIF_GNN_test and Dockground v1 datasets. Our method EquiRank outperforms all other competing methods by achieving the highest per-target Spearman on both VoroIF_GNN_test and Dockground v1 datasets. On VoroIF_GNN_test set EquiRank outperforms all other competing methods by attaining the highest Spearman correlation of 0.703. EquiRank demonstrates a performance improvement of approximately 9% compared to the second-best method, AlphaFold_Multimer (0.703 vs 0.640). Furthermore, it achieves over a 10% improvement compared to the latest GNN-based complex quality estimation method, VoroIF_GNN. EquiRank achieves a significantly higher Spearman correlation than other GNN-based methods, including DProQA and GNN-DOVE, as well as DeepRank, integrated with pLM embeddings, while substantially improving performance over our recent method, PIQLE. Additionally, on the Dockground v1 dataset, EquiRank attains the highest Spearman correlation of 0.473 than any other competing methods. It is important to note that, AlphaFold-multimer, having the second-best per-target Spearman correlation on VoroIF_GNN dataset, attains a much lower correlation of 0.196 than EquiRank on this Dockground v1 dataset. Although the competing methods demonstrate better-ranking performance on the relatively balanced VoroIF_GNN_test dataset, their suboptimal performance on the Dockground v1 dataset, lacking balance between correct and incorrect protein multimeric complexes, reveals limitations of their generalizability. As such, EquiRank strikes a good balance in terms of its ranking performance on both near-balance and imbalance datasets. Additionally, it is important to note that, EquiRank improves the ranking performance by almost 22% (0.473 vs 0.370) than the second-best TRScore, while improving the performance over our recent protein-protein interface quality estimation method PIQLE. Other latest GNN-based approaches including GDockScore, DPoQA, and GNN-DOVE consistently demonstrate sub-optimal performance, with GNN-DOVE having a negative per-target Spearman correlation of -0.047. Additionally, it is worth noting that EquiRank consistently demonstrates better performance than the latest pLM-based method, DeepRank-GNN-esm, indicating the collective contribution of both the Protein Language Model (pLM)-based embeddings and the symmetry-aware Equivariant Neural Network.

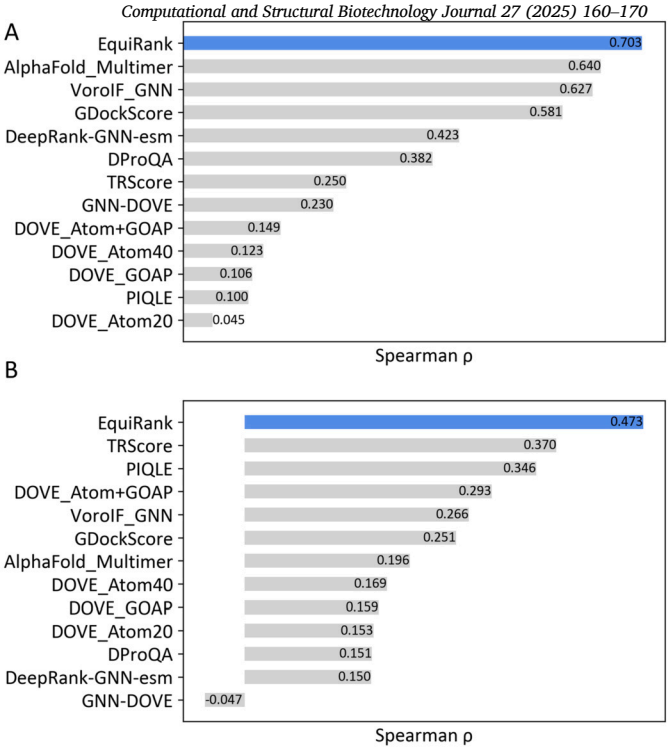


**Fig. 2.** Ranking performance of EquiRank and competing methods in terms of per-target average Spearman correlation coefficient ($\rho$) on A) VoroIF_GNN_test and B) DockGround v1 datasets.

We further evaluate methods' ranking performance in terms of the Top-N hit rate and success rate on VoroIF_GNN_test and Dockground v1 datasets as shown in Fig. 3. EquiRank consistently shows better hit rates compared to other competing methods starting from the top-10, while achieving comparable hit rates for top-1 and top-5 models. For instance, EquiRank achieves the highest top-10 hit rate of around 98% on VoroIF_GNN_test dataset (Fig. 3A) and around 43% on Dockground v1 (Fig. 3B) dataset. Notably, EquiRank significantly outperforms the language model-based method DeepRank-GNN-esm and improves the hit rate over our prior method PIQLE across both datasets. In terms of success rate, EquiRank achieves competitive Top-N success rates on both the VoroIF_GNN_test (Fig. 3C) and Dockground v1 (Fig. 3D) datasets. Specifically, EquiRank outperforms most of the competing methods for Top-1 model on VoroIF_GNN_test dataset (Fig. 3C). Additionally, EquiRank achieves a strong success rate close to 100% for Top-5 models, which is comparable to other high-performing methods such as VoroIF_GNN and GDockScore. On the Dockground v1 dataset (Fig. 3D), EquiRank achieves a stronger Top-1 success rate of 52% with VoroIF_GNN attaining the highest Top-1 success rate of 60%. However, EquiRank's success rate steadily improves as Top-N increases,

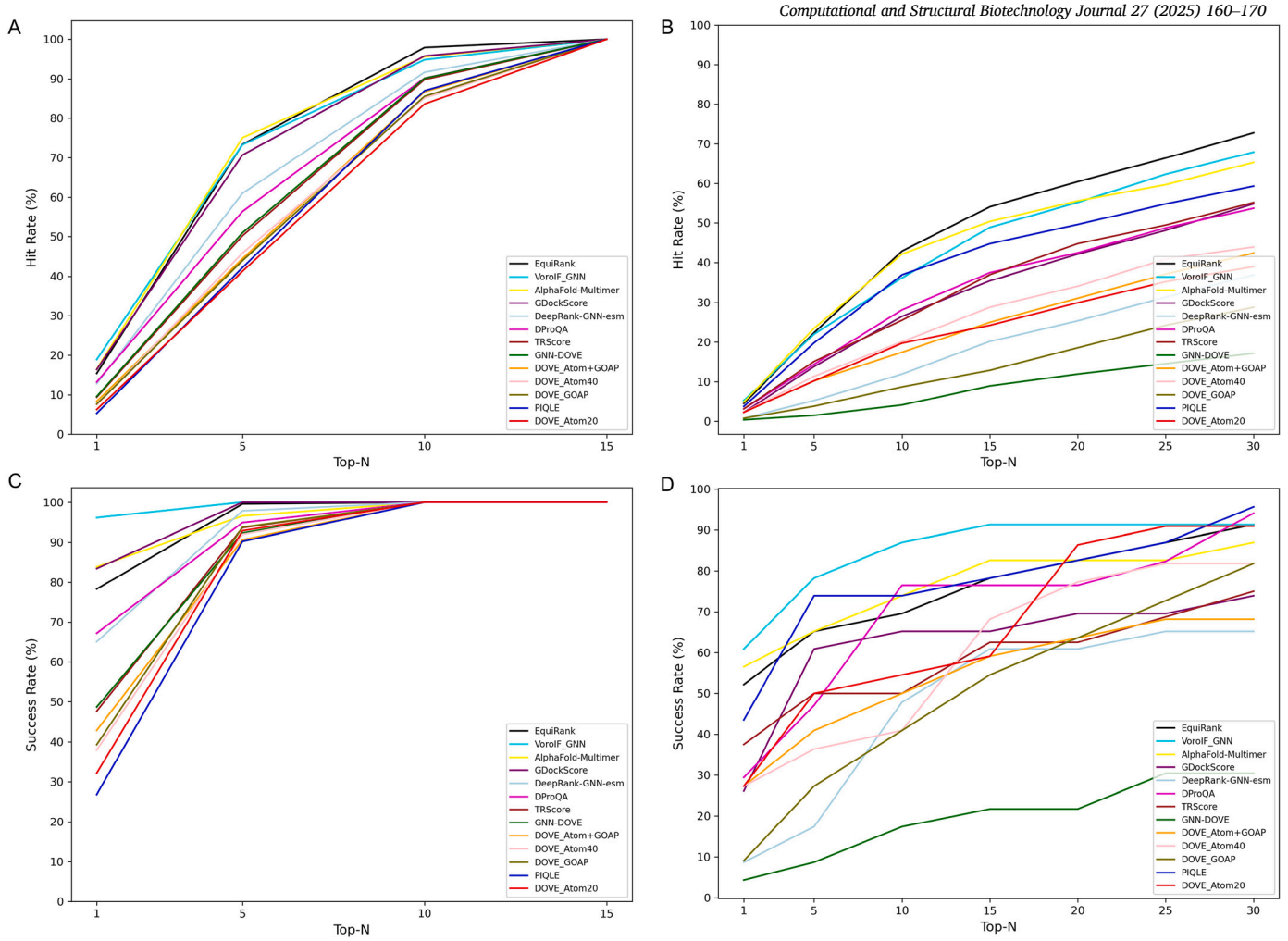


**Fig. 3.** Ranking complex structural models for EquiRank and the competing methods in terms of hit rate on (A) VoroIF_GNN_test dataset, (B) Dockground v1 dataset and success rate on (C) VoroIF_GNN_test dataset, (D) Dockground v1 dataset based on top-1, top-5, top-10, top-15 models for VoroIF_GNN_test and top-1, top-5, top-10, top-15, top-20, top-25 and top-30 models for Dockground v1 dataset. A DockQ threshold of 0.23 is used to identify acceptable models.

consistently outperforming other graph neural network-based methods, including GNN-DOVE, DProQA, and DeepRank-GNN-esm. Overall, EquiRank demonstrates improved ranking ability across a broad range of predicted protein complex models as demonstrated by various evaluation metrics.

### 3.2. Ability to distinguish high-quality models

Fig. 4 shows the ability of EquiRank and other competing methods method to successfully distinguish high-quality models on both Dockground v1 and CASP15 benchmark datasets. As shown in Fig. 4A, EquiRank attains the highest AUC of 0.968 on Dockground v1 dataset among all other competing methods. The GNN-based VoroIF_GNN attains the second-best AUC of 0.936. However, other GNN-based methods including GNN-DOVE, DProQA, and DeepRank-GNN-esm have limited distinguishability, having much lower AUC compared to EquiRank. On CASP15 datasets, as shown in Fig. 4B, EquiRank once again attains the highest AUC of 0.915 which is closely followed by the second-best performing method GDockScore with an AUC of 0.908 while also outperforming the Euclidean graph neural network based method EuDockScore (0.915 vs 0.824). Although VoroIF_GNN achieves the second-best AUC in Dockground v1, its performance is notably lower with an AUC of 0.849 compared to EquiRank.

DeepRank-GNN-esm, utilizing Protein Language Models (pLM)-based ESM-2 embeddings, demonstrates better performance in CASP15 (0.336 vs 0.835) than in Dockground v1, yet notably lower than EquiRank. It is noteworthy to mention that the CASP15 and Dockground v1 datasets exhibit significantly different balance ratios, with CASP15 being much more balanced than Dockground v1 in terms of correct and incorrect models as shown in Table 1. While the performance of other competing methods varies when applied to different datasets with diverse model qualities, EquiRank consistently achieves better performance, demonstrating its improved ability to distinguish high-quality complex models.

We additionally benchmark the distinguishability of EquiRank and other competing methods on both Dockground v1 and CASP15 datasets in terms of Area Under the Precision-Recall Curve (AUPRC). As shown in Fig. 5, while PIQLE achieves the highest PRAUC of 0.509 on the Dockground v1 dataset (Fig. 5A), EquiRank consistently achieves better PRAUC on both the Dockground v1 (Fig. 5A) and CASP15 (Fig. 5B) datasets, indicating its generalizability across diverse test datasets. On the Dockground V1 dataset (Fig. 5A), EquiRank achieves the second-best PRAUC of 0.335, which is comparatively lower than PIQLE's PRAUC of 0.509. This difference is due to the highly imbalanced nature of the Dockground v1 dataset, which has a lower percentage of correct decoys (Table 1). On the nearly balanced CASP15 dataset, EuDockScore, which incorporates natural protein language model-based embeddings, achieves the highest PRAUC of 0.455, indicating the contribution of the language model. EquiRank follows with the second-highest PRAUC of 0.326, outperforming all other competing methods, while other top-performing methods including GDockScore and AlphaFold-

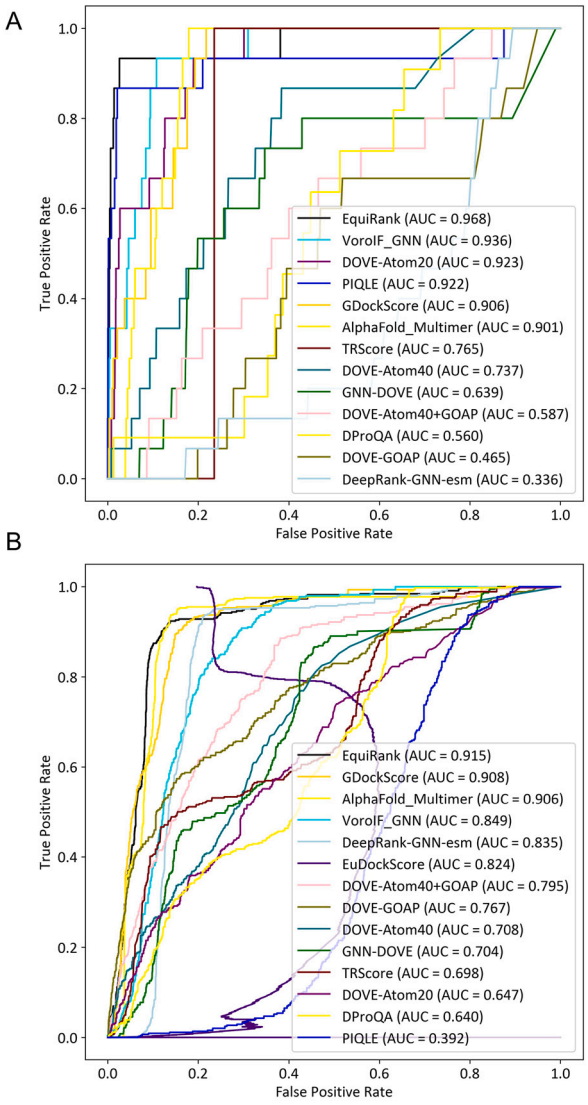


**Fig. 4.** Distinguishability of high-quality models for EquiRank and other competing methods, in terms of Area Under the Curve (AUC), on (A) Dockground V1 and (B) CASP15 datasets with a DockQ threshold of 0.8.

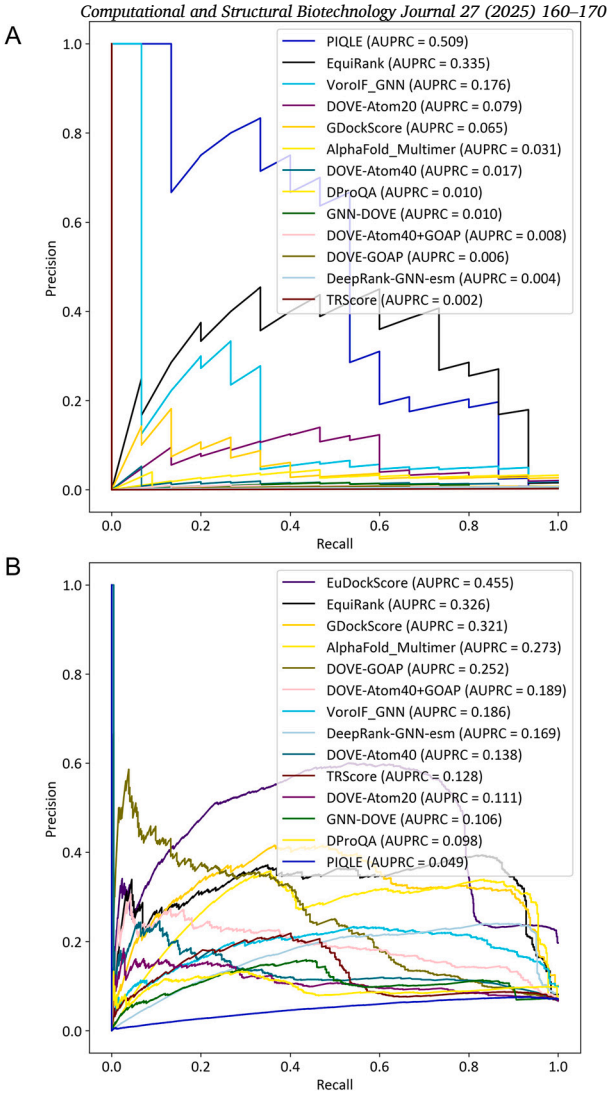


**Fig. 5.** Distinguishability of high-quality models for EquiRank and other competing methods, in terms of Area Under the Precision-Recall Curve (PRAUC), on (A) Dockground V1 and (B) CASP15 datasets with a DockQ threshold of 0.8.

Multimer achieve PRAUCs of 0.321 and 0.273, respectively. It is noteworthy that EquiRank demonstrates substantial performance improvement compared to other graph neural network-based methods using protein-language model embeddings, such as DeepRank-GNN-esm, on both the Dockground v1 (0.335 vs. 0.004) and CASP15 (0.326 vs. 0.169) datasets. Overall, EquiRank's consistent performance on both datasets highlights its generalizability and the effectiveness of its framework in distinguishing high-quality complex models.

### 3.3. Ablation study

To evaluate the relative importance of the features and network architecture used in EquiRank for ranking performance, we performed a series of feature and network ablations as shown in Fig. 6. Each feature ablation experiment involves isolating a single feature and training an ensemble of EGNN-based EquiRank models. Additionally, each network ablation involves isolating a network component from the EGNN or the EGNN network itself and subsequently training the modified ensemble EGNN or the alternative network on all the features. We evaluate the ranking performance on the VoroIF_GNN_validation dataset by measuring the per-target average Spearman correlation. As shown in Fig. 6A, EquiRank with all features achieved the highest Spearman correlation of 0.662, demonstrating the cumulative contributions of all features to enhanced ranking performance. Notably, the ablation of Protein Language Model (pLM)-based ESM-2 features ("No pLM embeddings") resulted in the lowest ranking performance, with an average Spearman correlation of 0.480. This indicates the significant contribution of the pLM features in improving the performance by nearly 28%. Similarly, the improved Multiple Sequence Alignment (MSA) encoding provided by AlphaFold2 ("No MSA encoding") contributes substantially, with a performance increase exceeding 24%. To confirm the significance of these contributions, we performed a t-test, which resulted in a p-value of 0 for both ablations, further highlighting their impact on improving performance. Additionally, the novel representations of multimeric geometry utilized in our recent PIQLE method [22] contribute to improved ranking performance ("No multimeric geometry," "No multimeric distance," and "No multimeric orientation"). Although their contribution is not as substantial as that of sequence-based features, these representations can complement the sequence-based features, representing a significant improvement over PIQLE [22].

We also evaluate the contribution of our ensemble EGNN network by performing network ablations while training the networks with all the features as shown in Fig. 6B. We first evaluate the contribution of the equivariance properties in EGNNs by making the models invariant, effectively disabling the equivariance update ("No equivariance"). This

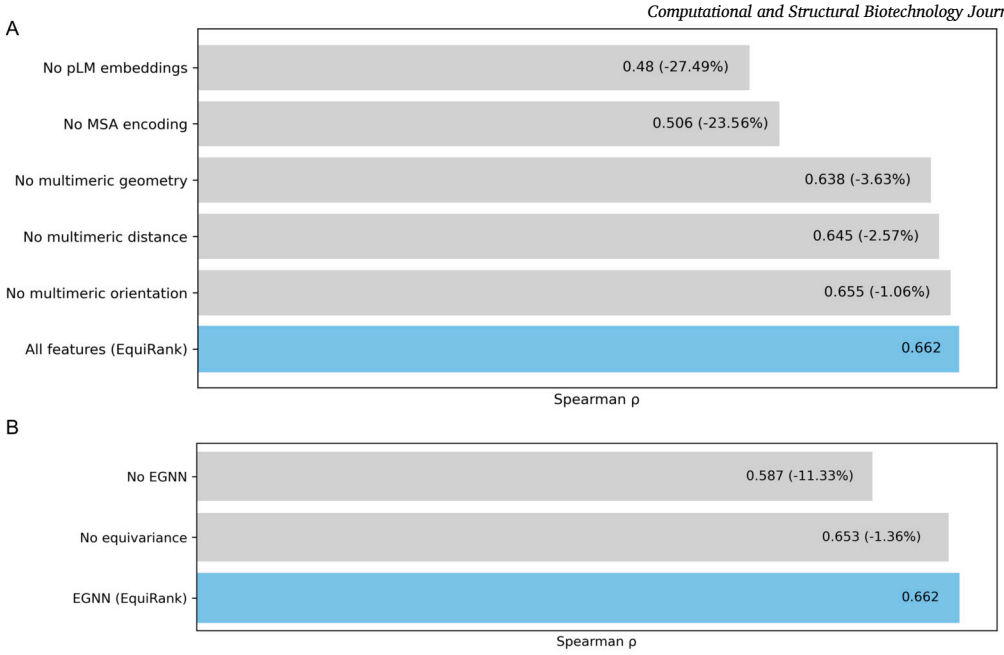


**Fig. 6.** Ablation study on the independent VoroIF_GNN_validation validation dataset in terms of per-target average Spearman correlations coefficient ($\rho$) between the estimated qualities of the protein-protein interfaces and their corresponding DockQ scores.

adjustment worsens the ranking performance, decreasing the per-target average correlation from 0.662 to 0.653. This suggests the relative importance of maintaining equivariance properties during the learning of protein-protein interface graphs. Additionally, to evaluate the contribution of our overall ensemble EGNN network, we train an ensemble of Graph Attention Networks (GATs) as used in our previous work, PIQLE [22], which was demonstrated to perform better than other Graph Neural Networks such as Graph Transformer Network and Graph Convolutional Neural Network. Fig. 6B shows that ensemble GATs ("No EGNN") worsen the performance by more than 11%, signifying the contribution of our ensemble EGNN network to improved protein-protein interface quality estimation performance. Once again to confirm the significance of this contribution, we performed a t-test, which resulted in a p-value of 0, underscoring the importance of the EGNN network in improving performance. Overall, our ablation studies underscore the significant contributions of both the features including Protein Language Model-based features and the equivariant neural network to the improvement of the protein-protein interface quality estimation performance of EquiRank.

## 4. Conclusions

In this work, we present EquiRank, an improved protein-protein interface quality estimation method. EquiRank introduces several new advances over our recent protein-protein interface quality estimation method PIQLE including the application of a symmetry-aware ensemble Equivariant Graph Neural Network and integration of several new sequence- and structure-based features including Protein Language Model-based ESM-2 embeddings. Through extensive benchmarking on a wide range of datasets covering experimentally determined heterodimeric biological assemblies, X-ray unbound benchmark set, and CASP15 targets, EquiRank consistently outperforms PIQLE and other state-of-the-art protein complex quality estimation methods. Additionally, while existing protein complex quality estimation methods show inconsistent performance, EquiRank's improved ability to rank protein complex models and distinguish high-quality models across datasets with varying distributions, from balanced to imbalanced, demonstrates its generalizability across large-scale benchmark datasets. Finally, through our ablation studies, we demonstrate that the improved performance is directly related to the application of equivariance properties of EGNN, and the integration of new sequence-based representation including protein language model-based embeddings.

## CRediT authorship contribution statement

**Md Hossain Shuvo:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Debswapna Bhattacharya:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work was partially supported by the National Institute of General Medical Sciences [R35GM138146 to D.B.] and the National Science Foundation [DBI2208679 to D.B.]. The authors gratefully acknowledge high-performance computing resources and technical support from the National AI Research Resource Pilot award (NAIRR240093 to D.B.).

## Data availability

The raw data used in this study, including the datasets for train, test, and validation as mentioned in Table 1 are collected from publicly available sources.

**VoroIF_GNN_train**, **VoroIF_GNN_test**, and **VoroIF_GNN_validation** datasets are available at https://dx.doi.org/10.5281/zenodo.7841307, **CASP13** train dataset is available at https://predictioncenter.org/download_area/CASP13/predictions/oligo/, **CASP14** train dataset is available at https://predictioncenter.org/download_area/CASP14/predictions/oligo/, **CASP15** test dataset is available at https://predictioncenter.org/download_area/CASP15/predictions/oligo/, **Dockground**

**v1** test dataset is available at https://dockground.compbio.ku.edu/downloads/unbound/decoy/decoys1.0.zip.

## References


[1] Peng X, Wang J, Peng W, Wu F-X, Pan Y. Protein-protein interactions: detection, reliability assessment and applications. Brief Bioinform 2017;18(5):798–819. https://doi.org/10.1093/bib/bbw066.

[2] Lu H, Zhou Q, He J, Jiang Z, Peng C, Tong R, et al. Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. Signal Transduct Targeted Ther 2020;5(1):1–23. https://doi.org/10.1038/s41392-020-00315-3. Publisher: Nature Publishing Group.

[3] Zaki N, Efimov D, Berengueres J. Protein complex detection using interaction reliability assessment and weighted clustering coefficient. BMC Bioinform 2013;14(1):163. https://doi.org/10.1186/1471-2105-14-163.

[4] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596(7873):583–9. https://doi.org/10.1038/s41586-021-03819-2. Publisher: Nature Publishing Group.

[5] Zahiri J, Emamjomeh A, Bagheri S, Ivazeh A, Mahdevar G, Sepasi Tehrani H, et al. Protein complex prediction: a survey. Genomics 2020;112(1):174–83. https://doi.org/10.1016/j.ygeno.2019.01.011. https://www.sciencedirect.com/science/article/pii/S088875431830572X.

[6] Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2. Nat Commun 2022;13(1):1265. https://doi.org/10.1038/s41467-022-28865-w. Publisher: Nature Publishing Group.

[7] Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer. pages: 2021.10.04.463034 Section: New Results (Mar. 2022). https://doi.org/10.1101/2021.10.04.463034. https://www.biorxiv.org/content/10.1101/2021.10.04.463034v2.

[8] Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold3. Nature May 2024. https://doi.org/10.1038/s41586-024-07487-w.

[9] Christoffer C, Bharadwaj V, Luu R, Kihara D. LZerD protein-protein docking webserver enhanced with de novo structure prediction. Front Mol Biosci 2021;8. https://www.frontiersin.org/articles/10.3389/fmolb.2021.724947.

[10] Pierce BG, Wiehe K, Hwang H, Kim B-H, Vreven T, Weng Z. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics 2014;30(12):1771–3. https://doi.org/10.1093/bioinformatics/btu097.

[11] Lyskov S, Gray JJ. The RosettaDock server for local protein-protein docking. Nucleic Acids Res 2008;36:233–8. https://doi.org/10.1093/nar/gkn216. Web Server issue.

[12] Gao M, Nakajima An D, Parks JM, Skolnick J. AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. Nat Commun 2022;13(1):1744. https://doi.org/10.1038/s41467-022-29394-2. Publisher: Nature Publishing Group.

[13] Sandor V, Kozakov D. Sampling and scoring: a marriage made in heaven. Proteins 2013;81(11):1874–84. https://doi.org/10.1002/prot.24343. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3942495/.

[14] Cao Y, Shen Y. Energy-based graph convolutional networks for scoring protein docking models. Proteins 2020;88(8):1091–9. https://doi.org/10.1002/prot.25888. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7374013/.

[15] Guo L, He J, Lin P, Huang S-Y, Wang J. TRScore: a 3D RepVGG-based scoring method for ranking protein docking models. Bioinformatics 2022;38(9):2444–51. https://doi.org/10.1093/bioinformatics/btac120.

[16] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv: 1512.03385, Dec. 2015. http://arxiv.org/abs/1512.03385.

[17] Wang X, Flannery ST, Kihara D. Protein docking model evaluation by graph neural networks. Front Mol Biosci 2021;8. https://www.frontiersin.org/article/10.3389/fmolb.2021.647915.

[18] O'Shea K, Nash R. An introduction to convolutional neural networks. arXiv:1511.08458 [cs], Dec. 2015. https://doi.org/10.48550/arXiv.1511.08458. http://arxiv.org/abs/1511.08458.

[19] Stebliankin V, Shirali A, Baral P, Shi J, Chapagain P, Mathee K, et al. Evaluating protein binding interfaces with transformer networks. Nat Mach Intell 2023;5(9):1042–53. https://doi.org/10.1038/s42256-023-00715-4. publisher Nature Publishing Group.

[20] Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, et al. Graph neural networks: a review of methods and applications. AI Open 2020;1:57–81. https://doi.org/10.1016/j.aiopen.2021.01.001. https://linkinghub.elsevier.com/retrieve/pii/S2666651021000012.

[21] Keramatfar A, Rafiee M, Amirkhani H. Graph neural networks: a bibliometrics overview. Mach Learn Appl 2022;10:100401. https://doi.org/10.1016/j.mlwa.2022.100401. https://www.sciencedirect.com/science/article/pii/S2666827022000780.

[22] Shuvo MH, Karim M, Roche R, Bhattacharya D. PIQLE: protein-protein interface quality estimation by deep graph learning of multimeric interaction geometries. Adv Bioinform 2023;3(1). vbad070. https://doi.org/10.1093/bioadv/vbad070.

[23] Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. arXiv:1710.10903, Feb. 2018. http://arxiv.org/abs/1710.10903.

[24] Olechnovic K, Venclovas Č. VoroIF-GNN: Voronoi tessellation-derived protein-protein interface assessment using a graph neural network. Proteins Jul. 2023. https://doi.org/10.1002/prot.26554.

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

[26] Chen X, Morehead A, Liu J, Cheng J. A gated graph transformer for protein complex structure quality assessment and its performance in CASP15. Bioinformatics 2023;39(1). i308–i317. https://doi.org/10.1093/bioinformatics/btad203.

[27] McFee M, Kim PM. GDockScore: a graph-based protein-protein docking scoring function. Adv Bioinform 2023;3(1). vbad072. https://doi.org/10.1093/bioadv/vbad072.

[28] Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 2022;44(10):7112–27. https://doi.org/10.1109/TPAMI.2021.3095381.

[29] Ferruz N, Schmidt S, Hocker B. ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun 2022;13(1):4348. https://doi.org/10.1038/s41467-022-32007-7. Publisher: Nature Publishing Group.

[30] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci 2021;118(15):e2016239118. https://doi.org/10.1073/pnas.2016239118. publisher: Proceedings of the National Academy of Sciences.

[31] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379(6637):1123–30. https://doi.org/10.1126/science.ade2574. publisher: American Association for the Advancement of Science.

[32] Chowdhury R, Bouatta N, Biswas S, Floristean C, Kharkar A, Roy K, et al. Single-sequence protein structure prediction using a language model and deep learning. Nat Biotechnol 2022;40(11):1617–23. https://doi.org/10.1038/s41587-022-01432-w. Publisher: Nature Publishing Group.

[33] Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, et al. Large language models generate functional protein sequences across diverse families. Nat Biotechnol 2023;41(8):1099–106. https://doi.org/10.1038/s41587-022-01618-2. Publisher: Nature Publishing Group.

[34] Horne J, Shukla D. Recent advances in machine learning variant effect prediction tools for protein engineering. Ind Eng Chem Res 2022;61(19):6235–45. https://doi.org/10.1021/acs.iecr.1c04943. publisher: American Chemical Society.

[35] Wu F, Wu L, Radev D, Xu J, Li SZ. Integration of pre-trained protein language models into geometric deep learning networks. Commun Biol 2023;6(1):1–8. https://doi.org/10.1038/s42003-023-05133-1. publisher: Nature Publishing Group.

[36] Liu D, Zhang B, Liu J, Li H, Song L, Zhang G. Assessing protein model quality based on deep graph coupled networks using protein language model. Brief Bioinform 2024;25(1):bbad420. https://doi.org/10.1093/bib/bbad420.

[37] Réau M, Renaud N, Xue LC, Bonvin AMJJ. DeepRank-GNN: a graph neural network framework to learn patterns in protein–protein interfaces. Bioinformatics 2023;39(1):btac759. https://doi.org/10.1093/bioinformatics/btac759.

[38] Yun S, Jeong M, Kim R, Kang J, Kim HJ. Graph transformer networks. Advances in neural information processing systems, vol. 32. Curran Associates, Inc.; 2019. https://proceedings.neurips.cc/paper/2019/hash/9d63484abb477c97640154d40595a3bb-Abstract.html.

[39] Yu H-X, Wu J, Yi L. Rotationally equivariant 3D object detection. IEEE Comput Soc 2022:1446–54. https://doi.org/10.1109/CVPR52688.2022.00151. https://www.computer.org/csdl/proceedings-article/cvpr/2022/694600b446/1H1k0VMzxfi.

[40] Satorras VG, Hoogeboom E, Welling M. E(n) equivariant graph neural networks. arXiv:2102.09844 [cs, stat], Feb. 2022. http://arxiv.org/abs/2102.09844.

[41] Chen C, Chen X, Morehead A, Wu T, Cheng J. 3D-equivariant graph neural networks for protein model quality assessment. Bioinformatics 2023;39(1):btad030. https://doi.org/10.1093/bioinformatics/btad030.

[42] Roche R, Moussad B, Shuvo MH, Tarafder S, Bhattacharya D. EquiPNAS: improved protein-nucleic acid binding site prediction using protein-language-model-informed equivariant deep graph neural networks. Nucleic Acids Res 2024;52(5):e27. https://doi.org/10.1093/nar/gkae039.

[43] Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods 2022;19(6):679–82. https://doi.org/10.1038/s41592-022-01488-1. Publisher: Nature Publishing Group.

[44] Steinegger M, Soding J. Clustering huge protein sequence sets in linear time. Nat Commun 2018;9(1):2542. https://doi.org/10.1038/s41467-018-04964-5.

[45] Pierce B, Weng Z. A combination of rescoring and refinement significantly improves protein docking performance. Proteins 2008;72(1):270–9. https://doi.org/10.1002/prot.21920.

[46] Li Y, Hu J, Zhang C, Yu D-J, Zhang Y. ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. Bioinformatics 2019;35(22):4647–55. https://doi.org/10.1093/bioinformatics/btz291. https://academic.oup.com/bioinformatics/article/35/22/4647/5487385.

[47] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577–637. https://doi.org/10.1002/bip.360221211.

[48] Li H, Hou J, Adhikari B, Lyu Q, Cheng J. Deep learning methods for protein torsion angle prediction. BMC Bioinform 2017;18(1):417. https://doi.org/10.1186/s12859-017-1834-2.

[49] Ballester PJ, Richards WG. Ultrafast shape recognition to search compound databases for similar molecular shapes. J Comput Chem 2007;28(10):1711–23. https://doi.org/10.1002/jcc.20681.

[50] Guo S-S, Liu J, Zhou X-G, Zhang G-J. DeepUMQA: ultrafast shape recognition-based protein model quality assessment using deep learning. Bioinformatics 2022;38(7):1895–903. https://doi.org/10.1093/bioinformatics/btac056.

[51] Jing X, Xu J. Fast and effective protein model refinement using deep graph neural networks. Nat Comput Sci 2021;1(7):462–9. https://doi.org/10.1038/s43588-021-00098-9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8939834/.

[52] Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol 2011;487:545–74. https://doi.org/10.1016/B978-0-12-381270-4.00019-6.

[53] Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66–93. https://doi.org/10.1016/S0076-6879(04)83004-0.

[54] Shuvo MH, Bhattacharya S, Bhattacharya D. QDeep: distance-based protein model quality estimation by residue-level ensemble error classifications using stacked deep residual neural networks. Bioinformatics 2020;36(Supplement_1). i285–i291. https://doi.org/10.1093/bioinformatics/btaa455.

[55] Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci USA 2020;117(3):1496–503. https://doi.org/10.1073/pnas.1914677117.

[56] Kundrotas PJ, Anishchenko I, Dauzhenka T, Kotthoff I, Mnevets D, Copeland MM, et al. Dockground: a comprehensive data resource for modeling of protein complexes. Protein Sci 2018;27(1):172–81. https://doi.org/10.1002/pro.3295.

[57] Basu S, Wallner B. DockQ: a quality measure for protein-protein docking models. PLoS ONE 2016;11(8):e0161879. https://doi.org/10.1371/journal.pone.0161879. publisher: Public Library of Science.

[58] Lensink MF, Wodak SJ. Docking, scoring, and affinity prediction in CAPRI. Proteins, Struct Funct Bioinform 2013;81(12):2082–95. https://doi.org/10.1002/prot.24428. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.24428. https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.24428.

[59] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: ICML; 2006.

[60] Wang X, Terashi G, Christoffer CW, Zhu M, Kihara D. Protein docking model evaluation by 3D deep convolutional neural networks. Bioinformatics 2020;36(7):2113–8. https://doi.org/10.1093/bioinformatics/btz870.

[61] Roney JP, Ovchinnikov S. State-of-the-art estimation of protein model accuracy using AlphaFold. Phys Rev Lett 2022;129(23):238101. https://doi.org/10.1103/PhysRevLett.129.238101. publisher: American Physical Society.

[62] Wang M, Zheng D, Ye Z, Gan Q, Li M, Song X, et al. Deep graph library: a graph-centric, highly-performant package for graph neural networks. arXiv:1909.01315 [cs, stat], Aug. 2020. http://arxiv.org/abs/1909.01315.

[63] Kingma DP, Ba J. Adam: a method for stochastic optimization. https://doi.org/10.48550/arXiv.1412.6980, Dec. 2014. https://arxiv.org/abs/1412.6980v9.