Computational and Structural Biotechnology Journal 27 (2025) 478–487

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research article

Physical-aware model accuracy estimation for protein complex using deep learning method

Haodong Wang, Meng Sun, Lei Xie, Dong Liu, Guijun Zhang *

College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

ARTICLE INFO

Keywords:
Estimation of model accuracy
Single-model method
Protein complex structure prediction

ABSTRACT

With the breakthrough of AlphaFold2 on monomers, the research focus of structure prediction has shifted to protein complexes, driving the continued development of new methods for multimer structure prediction. Therefore, it is crucial to accurately estimate quality scores for the multimer model independent of the used prediction methods. In this work, we propose a physical-aware deep learning method, DeepUMQA-PA, to evaluate the residue-wise quality of protein complex models. Given the input protein complex model, the residue-based contact area and orientation features were first constructed using Voronoi tessellation, representing the potential physical interactions and hydrophobic properties. Then, the relationship between local residues and the overall complex topology as well as the inter-residue evolutionary information are characterized by geometry-based features, protein language model embedding representation, and knowledge-based statistical potential features. Finally, these features are fed into a fused network architecture employing equivalent graph neural network and ResNet network to estimate residue-wise model accuracy. Experimental results on the CASP15 test set demonstrate that our method outperforms the state-of-the-art method DeepUMQA3 by 3.69 % and 3.49 % on Pearson and Spearman, respectively. Notably, our method achieved 16.8 % and 15.5 % improvement in Pearson and Spearman, respectively, for the evaluation of nanobody-antigens. In addition, DeepUMQA-PA achieved better MAE scores than AlphaFold-Multimer and AlphaFold3 self-assessment methods on 43 % and 50 % of the targets, respectively. All these results suggest that physical-aware information based on the area and orientation of atom-atom and atom-solvent contacts has the potential to capture sequence-structure-quality relationships of proteins, especially in the case of flexible proteins. The DeepUMQA-PA server is freely available at http://zhanglab-bioinf.com/DeepUMQA-PA/.

1. Introduction

Protein-protein complexes are central in many crucial biological and cellular processes, which makes their structural elucidation important. With the significant progress made by AlphaFold2 [1] in single-chain structure prediction, the prediction of structures for protein multimers has become the focus of research in the field. Since the structure of protein complexes is the key to understanding its function, methods such as AlphaFold-Multimer [2], DMFold-Multimer [3], AFsample [4], trRosettaX2 [5], and the recently released AlphaFold3[6] have been actively developed to predict the structure of multimers. Nonetheless, challenges remain in predicting structures with weak evolutionary signals, such as nanobody-antigen and antibody-antigen complexes [7]. The CASP15 results show that most successful prediction methods for protein multimers used modifications of the standard AlphaFold, including extensive sampling through variations on MSA construction, the use of multiple seeds, an increased number of recycles and extensive network dropout [7]. It also shows that, at least for now, scoring and ranking the accurate models from many decoys has become a fundamental strategy for improving the accuracy of protein multimers structure prediction. Not surprisingly, estimation of model accuracy (EMA) of multimeric structures has recently received much attention in the field and has been introduced into CASP15 as a new prediction category [8].

Generally, the EMA methods of complexes are divided into two categories: multi-model methods and single-model methods. Multi-model methods require multiple models as input and then evaluate the quality of the models using structural alignments and strategies, such as MULTICOM_qa [9], ModFOLDdock [10] and VoroIF-jury [11]. Single-model methods require only a single protein complex structure as input, and do not require additional information to predict model

* Corresponding author.
E-mail address: zgj@zjut.edu.cn (G. Zhang).

https://doi.org/10.1016/j.csbj.2025.01.017
Received 30 October 2024; Received in revised form 18 January 2025; Accepted 21 January 2025
Available online 22 January 2025

quality, such as VoroIF-GNN [12], DeepUMQA3 [13], et.al. Multi-model methods have significant advantages in specific scenarios such as Critical Assessment of Techniques for Protein Structure Prediction (CASP). However, they rely heavily on the quality of the model pool, which is closely related to the accuracy of structure prediction. By contrast, single-model EMA methods are not restricted by the model pool and can score as well as select models lightly and efficiently. We have observed that the single-model EMA methods outperform the multi-model method, and even the CASP-specific consensus method, on the accuracy estimation track of complex interface contact residues of CASP15 in 2022 [14]. In addition, the single-model EMA method can not only evaluate the accuracy independently of structure prediction methods, but also can be a key component of multi-model methods. Naturally, it has become a frontier and research hotspot in the field of protein structure model quality assessment.

Most single-model EMA methods mainly consider the geometric and evolutionary factors of protein structure models and utilize the deep learning network to reveal the relationship between structure features and model quality. In recent two years, our in-house developed DeepUMQA3 [13] extracts feature from three levels of overall geometric topology, intra-chain and inter-chain, and use the improved deep residual neural network to predict the accuracy of interface residues. DeepUMQA3 extends the USR feature of the DeepUMQA series to protein complexes. The overall USR and inter-monomer USR of the complexes are used to capture the topological relationships between the global and local structure, as well as the topological relationships of the interchain. In addition, a residual neural network coupled with triangle update and axial attention is employed to predict the lDDT (local distance difference test) of each residue and the accuracy of interface residues. AlphaFold-Multimer uses the Evoformer module to encode multiple sequence alignment and template information to reflect evolutionary information and decodes the structural coordinates and quality scores in the Structure module [2]. However, it is challenging for the above methods to characterize the solvent effects of the surrounding environment of the protein surface, which is a crucial driver for the protein folding problem [15] and protein-protein interactions [16]. It is worth noting that VoroIF-GNN predicts the contact area accuracy for the complex interface by building a graph to represent the local contact and solvent surface based on Voronoi tessellation [12,17,18]. Inspired by concepts of atom-atom and atom-solvent contact areas, given the strong correlation between surface areas and physical interactions [19], it is reasonable to assume that the orientation characteristics of the contact surface may contain the crucial information of the native structure interface. This hypothesis suggests that, based on geometric and evolutionary features, further considering physical-aware information (e.g., the solvent energy characterized by contact surface area and orientation) may reveal more intrinsic relationships between protein structure and model accuracy.

Based on our previously developed DeepUMQA3 protocol, this work proposes a single-model method, DeepUMQA-PA, which is used for model scoring and ranking of multimeric protein structure. We design physical-aware features based on residue contacts to capture the relationship of hydrophobicity and orientation of the interface, while combining topological and protein sequence embedding to describe geometric and evolutionary features. These representations are fed into an equivalent graph neural network (EGNN) [20] coupled with the invariant point attention mechanism (IPA) [1] and a ResNet network to predict the per-residue accuracy estimation for protein multimers. The test results show that the physical-aware, geometric topological, and evolutionary features are complementary, and the use of these features can significantly improve the performance of accuracy estimation for protein complexes.

2. Methods

DeepUMQA-PA is a single-model protein complex EMA method, which includes four main parts: data preparation, feature extraction, network architecture, and residue-wise plDDT scores [21]. The schematic diagram of the designed pipeline is illustrated in Fig. 1, and each part of the pipeline will be described in detail below.

2.1. Feature extraction

We extract four classes of features from an input protein complex structure: physical-aware contact features (i.e., contact surface area and contact surface-based orientation features), geometry-based features (i.e., ultrafast shape recognition and voxelization features), embedding features (i.e., sequence and structure embedding features), and knowledge-based statistical potential energy features (i.e., Rosetta energy), as shown in Fig. 1B. Details of all these features are available in Table S1 in the supplementary material.

2.1.1. Physical-aware contact area feature

In existing literature, most protein complex EMA methods use a distance threshold, such as 5 Å or 8 Å, between specific atoms (e.g., C_β and C_α) to define the concept of contact[22,23]. Although this way can effectively characterize the spatial relationships between atoms, it cannot fully reflect the solvent effects of the environment around the protein surface, which are closely related to the atom-atom interactions. Inspired by the concepts of atom-atom and atom-solvent contact area [12,16–18], we use residue-level contact areas and contact surface-based orientation features to represent the strength of physical interactions between protein surface regions and their surrounding solvents, which may provide a new perspective and effective way to evaluate the accuracy of complex structure models.

Given an input complex structure, we first calculate the interatomic contact surface based on the Voronoi tessellation algorithm [17,18]. The formula is defined as follows:

$$\sqrt{((x-x_1)^2+(y-y_1)^2+(z-z_1)^2)}-r_1 = \sqrt{((x-x_2)^2+(y-y_2)^2+(z-z_2)^2)}-r_2 \quad (1)$$

where r_1 and r_2 represent the van der Waals radius of atom a_i and atom a_j respectively, and the point (x, y, z) lies on the contact surface equidistant from the van der Waals spheres of the two atoms (Fig. 2a).

For any contact pair formed by atoms a_i and a_j, the contact area can be calculated by using the triangulation algorithm [24] and Voronota software (version: 1.27.3834) [12]. Furthermore, we computed residue-level contact area (RCA) by adding the relevant atom-level contact areas (ACA). Specifically, we masked the contacts between atoms within residues and accumulated the contact areas between atoms of different residues to obtain the contact area feature matrix (L^* L^*1), where L is the number of residues in the complex. In fact, the contact area matrix between residues is a sparse matrix which is not conducive to the training of neural networks. Therefore, we further merged the contact areas between residues into the feature of the total contact area of a residue with all the surrounding contact residues (L^*1). The contact area feature between residues and solvent (L^*1) was obtained by accumulating the atomic-level solvent-accessible areas. Formally, the formula is defined as follows:

$$ACA_{(a_i,a_j)} = \sum_{k=1}^{L} \left|\overrightarrow{O_kP_k} \times \overrightarrow{O_kQ_k}\right| \Big/ 2 \quad (2)$$

$$RCA_{(r_a,r_b)} = \sum_{m=1}^{N} ACA^m_{(a_i,a_j)} \quad (3)$$

where $ACA_{(a_i,a_j)}$ represents the contact area between atom a_i and atom a_j. O_k, P_k, Q_k are the vertices of the kth triangle of the contact surface obtained by triangulation, L represents the total number of triangles. $RCA_{(r_a,r_b)}$ represents the contact area between residue r_a and r_b. m denotes

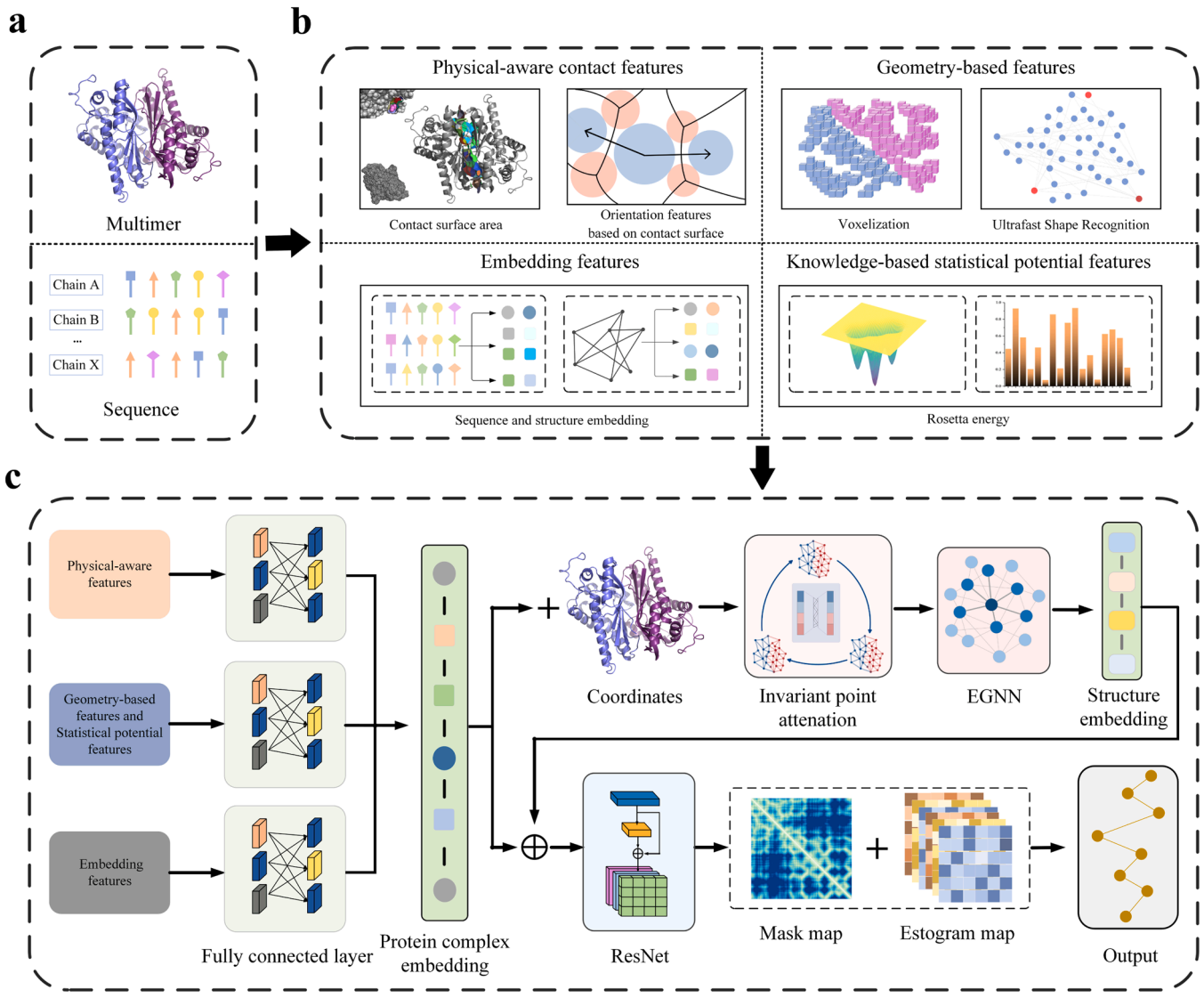

Fig. 1. The pipeline of DeepUMQA-PA. a) Data preparation. Protein complex structure and sequences were taken as input. b) Features extraction. Physical-aware contact features, geometry-based features, embedding features, and knowledge-based statistical energy features of given protein complex structure are extracted from the input structural and sequence information. c) Network architecture. The graph neural network was fused with ResNet network with attention mechanisms to estimate residue-wise prediction accuracy.

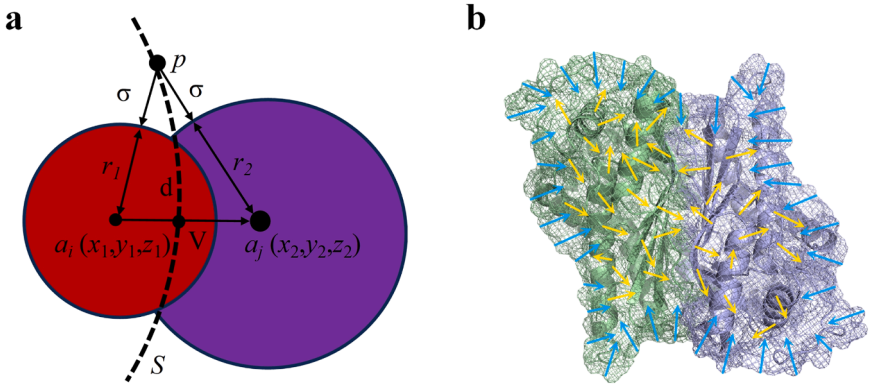

Fig. 2. a) Schematic diagram of atomic contact between atom a_i (red) and atom a_j (purple). r_1 and r_2 are the van der Waals radius of atom a_i and atom a_j, d is the distance between atoms, and p is a point on the Voronoi surface, where the surface distances from p to atom a_i and atom a_j are both σ. b) Schematic diagram of the residue-level contact surface orientation for homodimer 1B5D. The blue arrow represents the contact surface orientation of surface residues, and the yellow arrow represents the contact surface orientation of internal residues.

the mth atom-level contact area. N represents the total number of atom-level contact between residue r_a and r_b. The schematic diagrams of the residue contact surface and residue-solvent contact surface are shown in Fig. 3. The Voronota software (version: 1.27.3834) [12] and PyMOL (Version: 2.6.0a0) [25] were used for drawing Fig. 2 and Fig. 3.

2.1.2. Physical-aware contact orientation feature

To characterize the relationship between protein surface residues and internal residues, we designed residue-level contact surface orientation features, which is the sum of vectors of the contact orientation between a residue and all surrounding contact residues. First, for any two contact atoms $a_i\,(x_1, y_1, z_1)$ and $a_j\,(x_2, y_2, z_2)$ in a protein, we obtain the contact surface between atoms according to Voronoi tessellation [17]. The formula (1) in 2.1.1 can be simplified as follows:

$$\left(x-\frac{d}{2}\right)^2-\frac{\triangle r^2}{d^2-\triangle r^2}\left(y^2+z^2\right)=\frac{\triangle r^2}{4} \tag{4}$$

where $\triangle r$ is the difference in van der Waals radius between atom a_i and atom a_j, d is the distance between atoms.

Next, the contact surface S between atom a_i and atom a_j intersects the connecting line of the two atoms to obtain the reference point V (Fig. 2a). We use the support vector $\overrightarrow{N}_{\text{surface-vertex}(a_i,a_j)}$ at point V as the orientation of the interatomic contact surface. For two residues r_a and r_b, the residue-level contact surface orientation $\overrightarrow{N}_{\text{surface-vertex}(r_a,r_b)}$ is obtained by summing the support vectors of the relevant atom-level contact surface (Fig. 2b). Formally, the calculation formula is as follows:

$$\overrightarrow{N}_{\text{surface-vertex}\left(a_i,a_j\right)}=\left[2\left(x_v^{ij}-\frac{d}{2}\right),-\frac{2\triangle r^2}{d^2-\triangle r^2}y_v^{ij},-\frac{2\triangle r^2}{d^2-\triangle r^2}z_v^{ij}\right] \tag{5}$$

$$\overrightarrow{N}_{\text{surface-vertex}(r_a,r_b)}=\left[\sum_{m=1}^{n}2\left(x_v^{ij}-\frac{d}{2}\right),\sum_{m=1}^{n}\left(-\frac{2\triangle r_m^2}{d^2-\triangle r_m^2}y_v^{ij}\right),\sum_{m=1}^{n}\left(-\frac{2\triangle r_m^2}{d^2-\triangle r_m^2}z_v^{ij}\right)\right] \tag{6}$$

where $\left(x_v^{ij},y_v^{ij},z_v^{ij}\right)$ is the coordinate of reference point V, n represents the total number of atom-level contact surface between residues r_a and r_b, $\triangle r_m$ is the van der Waals radius difference of the m-th atom-level contact surface. More calculation formulas are listed in Text S1 in the supplementary material.

2.1.3. Geometry-based Voxelization and Ultrafast Shape Recognition (USR) features

For a protein complex, small conformational changes in local residues could cause a significant impact on inter-chain interactions and the overall structure. The voxelization features project the protein structure into voxelized grids to describe the local structural information of residues [26,27] while USR [13,28–31] can quickly capture the topological information of protein structures by using three sets of interatomic distances. Thereby, we use voxelization and overall USR to complementarily characterize the geometric topological relationship between local residues and the overall structure.

2.1.4. Embedding features and statistical potential features

Large Language Models (LLM) capture the evolutionary conservation information of proteins and have been widely used in protein structure prediction, design and function research [32]. The protein language model ESM can quickly and accurately obtain embedding information for structure and sequence [33]. Thereby, we use the high-dimensional embedding of protein structure of backbone atoms in ESM-IF1 [34] (1*L*512) and the high-dimensional embedding of sequence in ESM2 [35] (1*L*1280) to establish the connection between evolutionary conservation information and structural accuracy. In addition, Rosetta energy [36–38] features are also an important part of the input features, using the one-body-terms, the two-body energy terms and the presence of backbone-to-backbone hydrogen bonds features. All these features are normalized and fed to the deep learning neural network.

2.2. Network architecture

In this study, we designed a fusion network architecture, where each part is drawn from an equivariant graph neural network (EGNN) [20] coupled with invariant point attention (IPA) [1] and a ResNet network [39] with attention mechanism. The purpose of introducing EGNN network is to process the input structural coordinates, so that we can obtain a structure embedding representation that is closer to the native

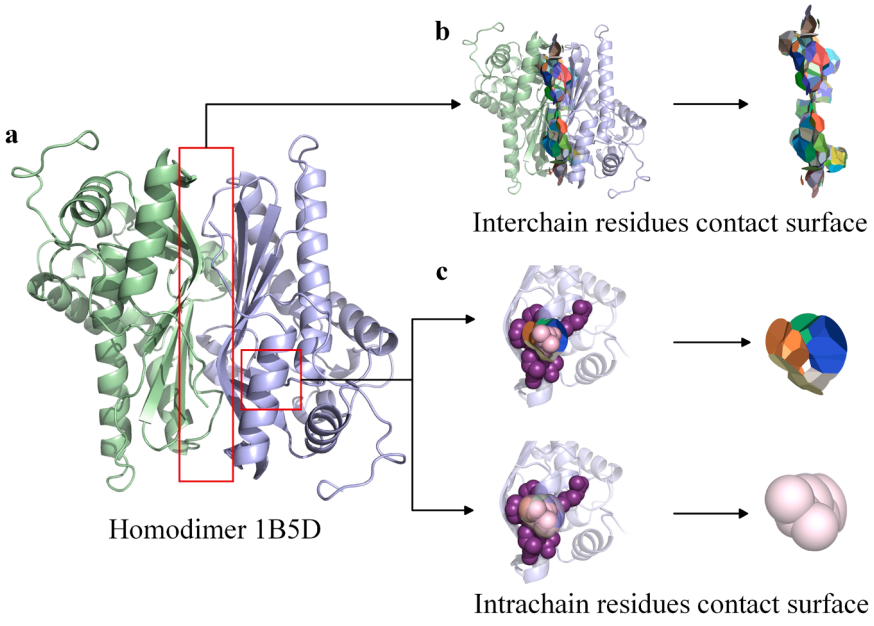

Fig. 3. a) Homodimer 1B5D. b) Interchain residues contact surface. Random color patches represent the contact surface of different contact residues from interchain. c) Intrachain residues contact surface. Random color patches represent the residue contact surface between the pink residue and the other contact residues (purple). Light pink patches represent the residue-solvent surface.

structure of protein. In this way, by combining feature embeddings from fully connected layers and input into the ResNet network as previously used in DeepUMQA3, we can make more accurate predictions on the distance contact (i.e., Mask) map and the distance error (i.e., Estogram) maps, and finally improve the performance of residue-wise accuracy estimation. The fusion network architecture is shown in Fig. 1C.

For a given protein complex structure model, we first extract four classes of protein features, which were physical-aware contact features, geometry-based features, embedding features, and knowledge-based statistical potential features. These features are firstly input into the fully connected layer [40] to generate a 1*L*128 protein complex embedding, which is combined with the coordinates of complex model and input into the EGNN network coupled with IPA module. EGNN network is used to iteratively update the protein atomic coordinates to better approximate the native structure. Then, output of EGNN (i.e., structure embedding) is recombined with the protein complex embedding and input into the residual network with attention mechanism [41], which consists of a main residual block and two branch residual blocks. Each residual block contains three 2D convolutional layers with different dilation rates [13], normalization layers and GELU activation function [42]. Finally, we get the mask map thresholded at 15 Å and estogram map with C_β distance deviations to calculate the residue-wise plDDT score by different branch residual blocks.

2.3. Training

We used a non-redundant protein complex data set from the DeepUMQA3 as the training and validation data (before January 1, 2022) and all the CASP15 data (as test data) we use are after May 2022. In this way, it can be ensured that there is no overlap between the training set and the test set and this allows for an objective and fair comparison with DeepUMQA3. The training and validation datasets contain a total of 7590 targets, each generating approximately 240 models. The ratio of the number of targets in the training and validation datasets is 9:1. Our best network model took 125 h to train on a single A100. The network Adam optimizer [43] with a learning rate of 0.01 is used, which decays at a rate of 0.05 %. During the training process of the network, the performance of the model is optimized by minimizing loss functions. Specifically, the cross-entropy loss function is used to evaluate the estogram map loss; the binary cross entropy loss function is used to evaluate the mask map loss; the root mean square deviation[44] is used to evaluate the coordinate loss and residue-wise plDDT score loss. Loss function is defined as follows:

$$Loss = w\left(L_{mas} + L_{est}\right) + L_{coor} + L_{plddt} \tag{7}$$

where w is the weight that is equal to 0.1, $L_{mas}$is the mask map loss, $L_{est}$is the estogram map loss, $L_{plddt}$is the residue-wise plDDT score loss, $L_{coor}$is the square value of the difference between the predicted and true protein structure coordinates during training. Specifically, during the training, the model coordinates and protein complex embedding features are input as node features and edge features into the EGNN network coupled with the IPA module to predict the true structure. Then, the square of the distance errors between the true structure coordinates and the predicted coordinates is used as the coordinate loss, which can obtain an embedding that approximates the true structure to guide the improvement of the model accuracy estimation. lDDT is used to analyze the stability of local regions of proteins. The calculation formula of lDDT is as follows [21]:

$$p_{1,2,3,4} = \frac{N_{\left(\left|D_{ij} - d_{ij}\right| < t\right)}}{N_{\left(D_{ij} < 15\right)}}, t \in \{0.5, 1, 2, 4\} \tag{8}$$

$$\mathrm{lDDT}_i = \frac{p_1 + p_2 + p_3 + p_4}{4p_0} \tag{9}$$

whereDijrepresents the distance between residue i and residue j in the reference structure, dijrepresents the distance between residue i and residue j in the prediction model. $N_{(|D_{ij}-d_{ij}|<t)}$represents the number of residues with the predicted distance error less than the threshold t Å, $t \in \{0.5, 1, 2, 4\}$. The distance error of residues is defined by the distance error of C_β.$N_{(D_{ij}<15)}$represents the number of residues that are within 15 Å from the reference structure.p_1, p_2, p_3, p_4respectively represent the probability of residues for which the absolute value of $|D_{ij} - d_{ij}| < t$Å, $t \in \{0.5, 1, 2, 4\}$. $p_0$represents the probability of residues within 15 Å of the residue i in the reference structure. The value range of lDDT is from 0 to 1. The closer the score is to 1, the closer the prediction model is to the reference structure.

3. Results and discussions

In this study, four statistical metrics are used to objectively and fairly analyze the reliability of predictions in complex interface residues. Specifically, we use Pearson[45] and Spearman [46] metrics to measure the correlation between the predicted lDDT score of interface residues and the true lDDT, focusing on evaluating the accuracy of the methods in ranking the models. ROC (AUC) [47] is used to evaluate the ability of the method to distinguish between high-quality and low-quality models. Regarding the definition of high-quality and low-quality models, we use the same definition as the official CASP definition [14]. Specifically, for all models of the same protein target, the models whose true quality is in the top 25 % are defined as high-quality models, and the other models are defined as low-quality models. The mean absolute error (MAE) quantifies the deviation between the predicted lDDT scores and the actual values.

3.1. Results on the CASP15 test set

We test the performance of DeepUMQA-PA in evaluating the accuracy of interface residues on the CASP15 test set. We used the interface threshold defined by the CASP specification, i.e, interface residues for protein complexes are defined as those with a C_β–C_β distance$\leq$ 8 Å between any two chains (or Cα in the case of glycine) [14]. Due to hardware resource limitations, we compare the performance of the state-of-the-art methods for evaluating local interface accuracy on 7875 models of 30 targets. We find that DeepUMQA-PA has advantages in the reliability of protein complex model scoring and ranking (Fig. 4). The detailed evaluation results are listed in Table S2 of the supplementary material. On average, DeepUMQA-PA outperforms other methods and improves over the top-performing method DeepUMQA3 by 3.69 %, 3.49 % and 0.48 % on Pearson, Spearman and ROC (AUC) of lDDT, respectively (Table 1). In particular, for the five nanobody-antigens targets (H1140-H1144), DeepUMQA-PA significantly improved by 16.8 %, 15.5 % and 5.1 % compared with DeepUMQA3 in the three statistical metrics of Pearson, Spearman and ROC (AUC) based on lDDT, respectively (Table 2). The interaction between antibody and antigen is formed through spatial structural complementarity. The smaller the distance between the two, the greater the interaction force (such as van der Waals force) [48,49]. This suggests that DeepUMQA-PA may have the potential to accurately assess the local geometry of protein binding sites [50]. This may be attributed to the fact that the introduction of physical-aware features enhances the ability of network to learn specific protein-protein interaction patterns.

3.2. Ablation studies

In the ablation study, we investigate the impact of physical-aware contact features on the performance of DeepUMQA-PA (Fig. 5). Specifically, we use the same training process to train multiple neural network models with different physical-aware features (i.e., physical-aware orientation, residue-residue contact area and residue-solvent contact

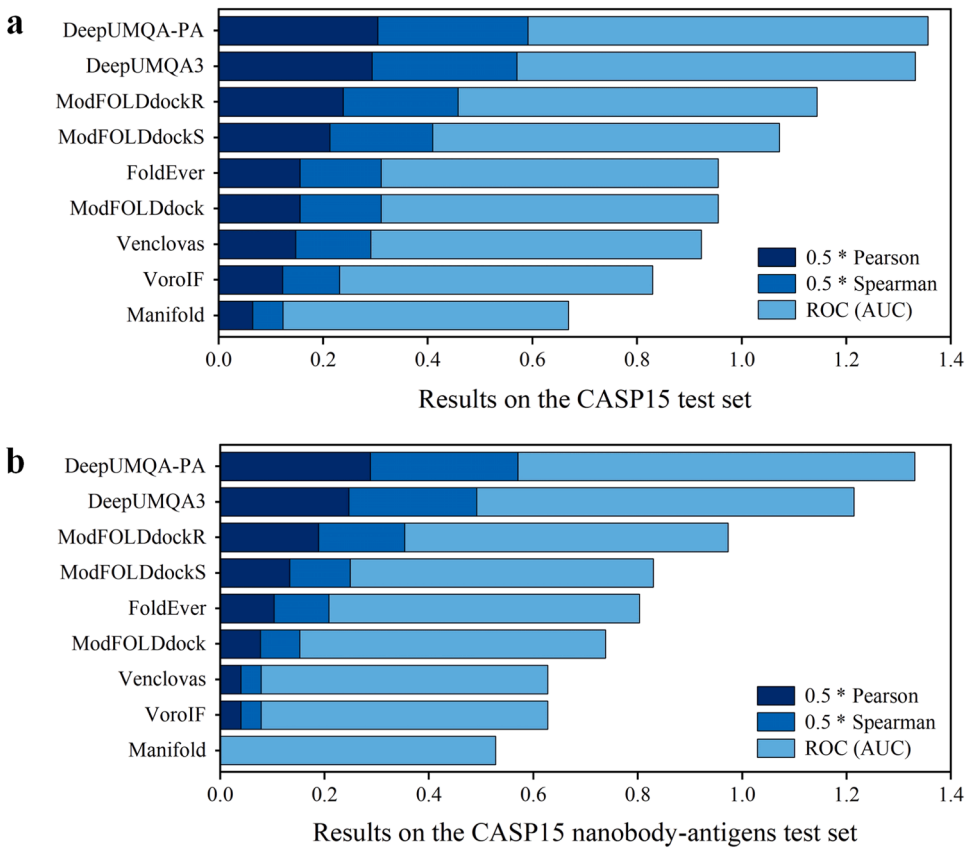

Fig. 4. **a** The results of DeepUMQA-PA and other servers on CASP15 test set. **b** The results of DeepUMQA-PA and other servers on CASP15 nanobody-antigens test set.

Table 1
Comparison of the evaluation results of Pearson, Spearman, ROC (AUC) on the CASP15 test set with other methods.

methods	Pearson	Spearman	ROC (AUC)
DeepUMQA-PA	**0.608**	**0.574**	**0.765**
DeepUMQA3	0.587	0.554	0.761
ModFOLDdockR	0.476	0.440	0.686
ModFOLDdockS	0.425	0.393	0.663
Venclovas	0.311	0.311	0.645
VoroIF	0.311	0.311	0.645
FoldEver	0.294	0.288	0.631
ModFOLDdock	0.245	0.218	0.598
Manifold	0.130	0.115	0.546

area) on the DeepUMQA3 data set, and use 29 protein complex targets from CASP15 as the test set. When the physical-aware contact orientation feature is removed from the baseline model DeepUMQA-PA (the full information is used), the Pearson and Spearman of lDDT decrease by 2.66 % and 1.80 %, respectively. In particular, for the five nanobody-antigen complexes, the Spearman and ROC (AUC) of lDDT decrease by 1.99 % and 1.46 %, respectively. The difference implies that the introduction of orientation information between residues enables DeepUMQA-PA to consider protein-protein docking and physical interaction mechanisms, which is crucial for identifying binding sites.

When the physical-aware residue-residue contact area feature is removed from the baseline model DeepUMQA-PA (the full information is used), the results show that the performance of DeepUMQA-PA significantly decreases by 9.72 %, 10.03 % and 3.29 % on the Pearson, Spearman and ROC (AUC) of lDDT respectively. This suggests that the residue-residue contact area feature may play a crucial role in characterizing the physical interactions between interface residues. When the residue-solvent contact area is not used, the performance of the model decreases most significantly, with the Pearson correlation decreasing from 0.61 to 0.52 and the Spearman correlation decreasing from 0.57 to 0.47. We also conduct tests on five nanobody-antigen complexes. Notably, the removal of the residue-solvent contact area feature results in a more obvious performance decrease (Pearson: 0.57–0.42, Spearman: 0.56–0.39, ROC AUC: 0.75–0.67). This is mainly because the distribution of solvent molecules near the antibody-antigen binding site affects the energy state of binding, and the residue-solvent contact area can reflect the solvent effects of the environment around the protein surface. In conclusion, the introduction of physical-aware contact features helps DeepUMQA-PA to accurately predict the quality of interface residues.

3.3. Comparison with AlphaFold-Multimer and AlphaFold3 self-estimation methods

AlphaFold-Multimer (AFM) and AlphaFold3 (AF3) can not only

Table 2
Comparison of the evaluation results of Pearson, Spearman, ROC (AUC) on the nanobody-antigen test set with other methods.

methods	Pearson	Spearman	ROC (AUC)
DeepUMQA-PA	**0.576**	**0.565**	**0.760**
DeepUMQA3	0.493	0.489	0.723
ModFOLDdockR	0.377	0.330	0.619
ModFOLDdockS	0.267	0.231	0.581
FoldEver	0.206	0.210	0.595
ModFOLDdock	0.154	0.150	0.586
Venclovas	0.080	0.078	0.549
VoroIF	0.080	0.078	0.549
Manifold	−0.152	−0.153	0.528

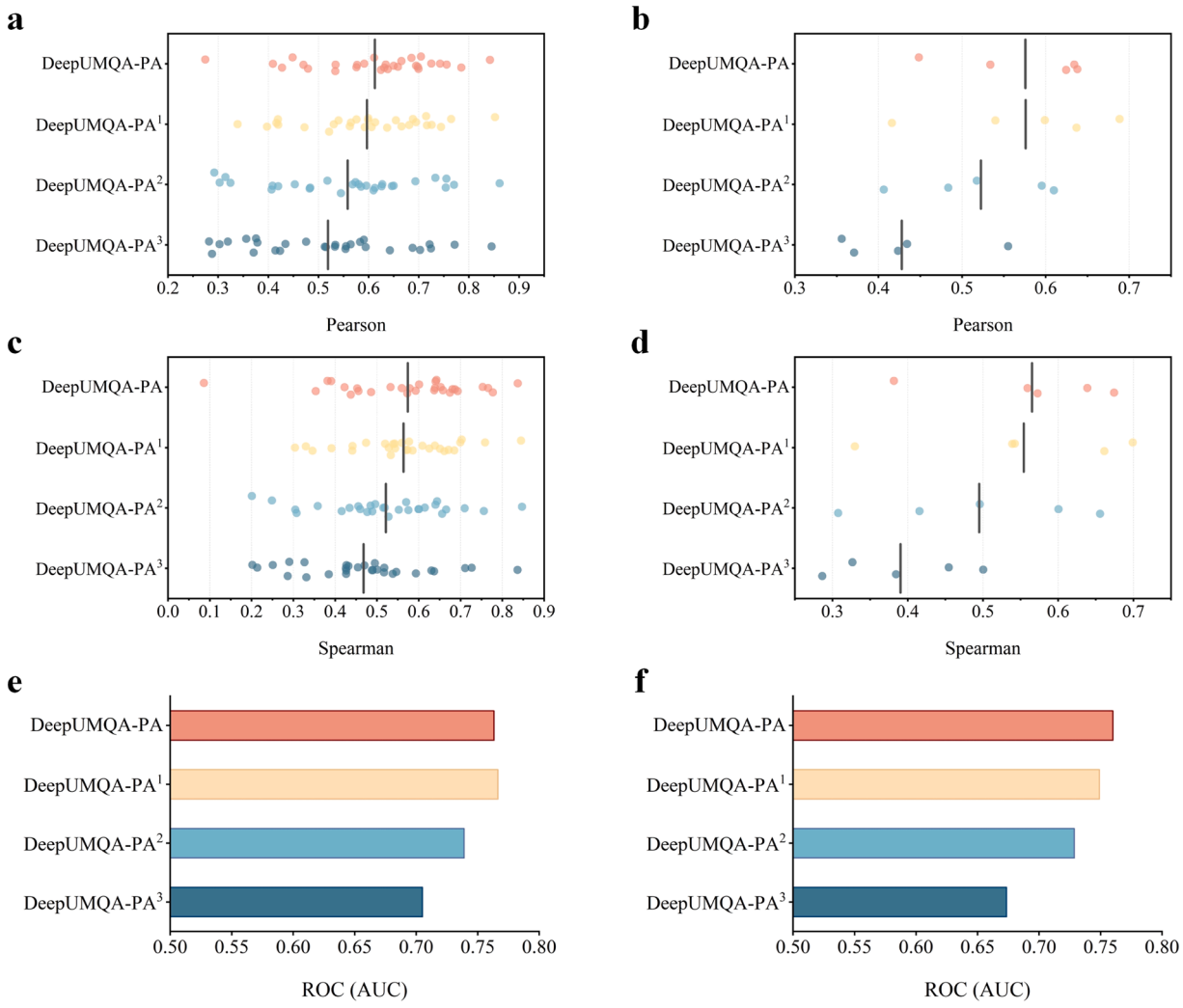

Fig. 5. a, c, e) Supplementary ablation studies of interface residue assessment accuracy on the CASP15 test set. b, d, f) Supplementary ablation studies of interface residue assessment accuracy on CASP15 nanobody-antigens test set. DeepUMQA-PA is the baseline model that uses all information. DeepUMQA-PA^1 denotes a version of DeepUMQA-PA without the physical-aware contact orientation feature. DeepUMQA-PA^2 denotes a version of DeepUMQA-PA without the physical-aware residue-residue contact area feature. DeepUMQA-PA^3 denotes a version of DeepUMQA-PA without the physical-aware residue-solvent contact area feature.

predict high-precision models but also provide reliable residue-wise confidence estimates. However, for the AF (i.e., AFM and AF3) model, there is still room for improvement in the accuracy of local evaluation. To analyze whether DeepUMQA-PA has the potential to improve evaluation accuracy for AF models with low self-assessment accuracy, we download the AFM prediction models provided by CASP15 and use the AF3 server to generate five models for each target. On average, AFM achieves essentially the same prediction as DeepUMQA-PA on MAE of lDDT for 30 targets (Table S3a). Fig. 6a shows the evaluation error comparison between DeepUMQA-PA and AFM on each target. DeepUMQA-PA achieves lower MAE scores than AFM on 43 % targets. We further find that DeepUMQA-PA improves more obviously on targets with a higher average MAE predicted by AFM. Especially for the target with MAE $>$ 0.1 between AFM plDDT scores and the true values, the average MAE of DeepUMQA-PA is significantly lower than that of AFM, with the average decrease from 0.169 to 0.140. Similarly, DeepUMQA-PA outperforms AF3 on 50 % targets, and improves by 15.17 % on these targets (Fig. 6b). These results highlight the synergic intersection between DeepUMQA-PA and AF in assessing local structural accuracy. DeepUMQA-PA provides complementary assessments in region with high uncertainty in AF predictions, which helps to accurately identify low-confidence regions of the prediction model to guide model refinement.

3.4. *Differences in quality estimation accuracy according to residue location*

We divide the protein complex model into core, interface, and surface residues, and observe differences in quality estimation accuracy according to residue location. The classification criteria for residue positions in the complex are as follows: (1) interface residues: C_β distance between chains$\leq$ 8 Å (Cα for glycine), (2)core residues: relative solvent-accessible surface area (SASA)$\leq$ 0.25, (3)surface residues: relative SASA $>$ 0.25.

We present the accuracy errors of core, surface, and interface residues predicted by DeepUMQA-PA in Fig. 7. On average, the evaluation accuracy of core residues is 24.79 %, 29.33 % higher than that of surface and interface residues, respectively (Table S4). Especially for the nanobody antigens (H1140-H1144), the gap in the MAE of lDDT between interface and core residues is even more obvious (core:0.074, surface:0.108, interface:0.174). This may be attributed to the fact that core residues are mainly used to maintain the overall stability of the protein, and their structural patterns are relatively simple and easy to evaluate. In contrast, the structures corresponding to surface and

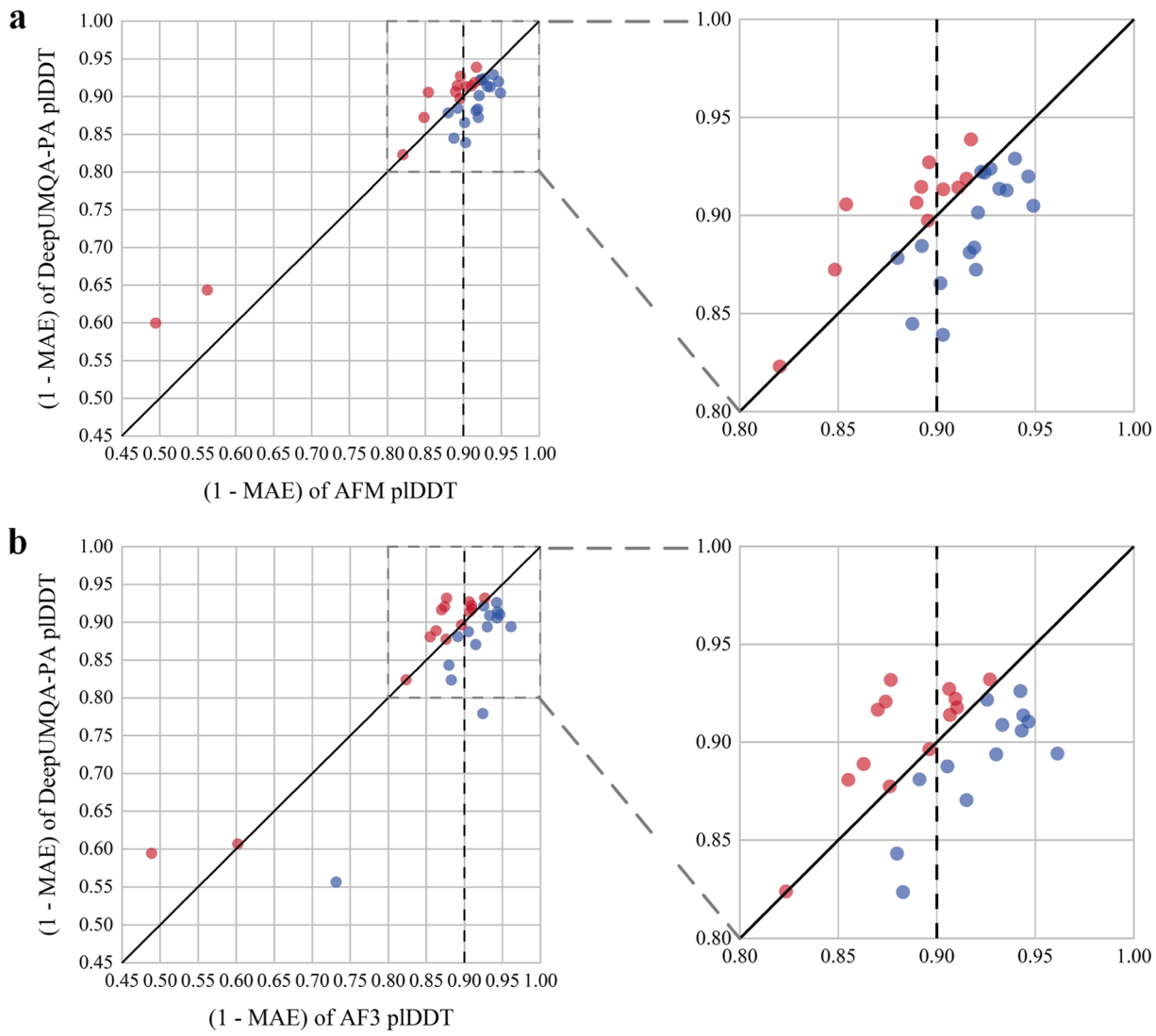

Fig. 6. a) Comparison of DeepUMQA-PA and AlphaFold - Multimer self-assessment accuracy in the mean absolute error (MAE). The models of 30 targets are generated by AFM. b) Comparison of DeepUMQA-PA and AlphaFold3 self-assessment accuracy in the mean absolute error (MAE). The models of 30 targets are generated by AF3 webserver.

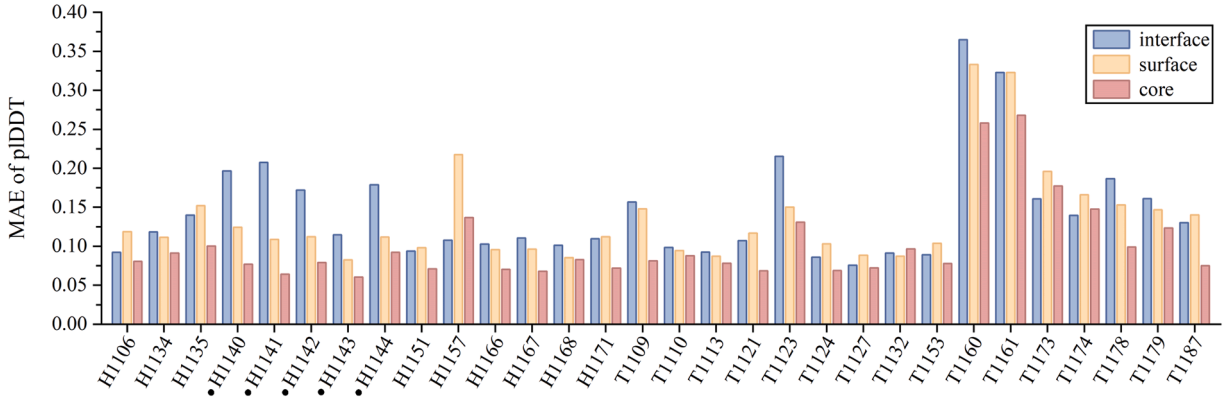

Fig. 7. Differences in quality estimation accuracy according to residue location in CASP15 targets. Blue represents interface residues, yellow represents surface residues, and red represents core residues.

interface residues may undergo conformational changes due to the influence of the surrounding environment or ligands, making the evaluation of these residues more challenging.

4. Conclusion

We developed a single-model EMA method for protein complex called DeepUMQA-PA. Based on DeepUMQA3, we further used physical-aware contact surface features (i.e. contact surface area and contact surface-based orientation features) and a fusion network architecture to evaluate the residue-wise model quality. Experimental results demonstrate that our method outperforms state-of-the-art EMA methods, including DeepUMQA3, ModFOLDdockR, ModFOLDdockS, VoroIF, Venclovas, FoldEver, ModFOLDdock and Manifold on 30 protein complex targets in CASP15. Ablation results demonstrate that physical-aware contact surface features can improve the performance of model quality assessment methods. In addition, for the MAE metric, our method is complementary to AlphaFold-Multimer and AlphaFold 3 in terms of local assessment accuracy and has an advantage over it in evaluating low-accuracy models. We further find that it is more challenging for DeepUMQA-PA to evaluate interface residues than core residues and surface residues. With the rapid development of complex structure prediction, model evaluation of protein binding to DNA, RNA and small molecules may be a research hotspot in the future.

CRediT authorship contribution statement

Xie Lei: Writing – review & editing, Visualization. **Liu Dong:** Writing – review & editing, Conceptualization. **Wang Haodong:** Writing – review & editing, Writing – original draft, Conceptualization. **Sun Meng:** Writing – review & editing, Writing – original draft, Conceptualization. **Zhang Guijun:** 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 no competing interests.

Acknowledgement

This study is supported by the National Key R & D Program of China (2022ZD0115103), the National Nature Science Foundation of China (62173304), Zhejiang Provincial Special Support Program for High-Level Talents (2023R5248).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2025.01.017.

References

[1] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9. https://doi.org/10.1038/s41586-021-03819-2.
[2] Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, et al. Protein complex prediction with AlphaFold-multimer. bioRxiv 2021. https://doi.org/10.1101/2021.10.04.463034.
[3] Zheng W, Wuyun Q, Freddolino PL, Zhang Y. Integrating deep learning, threading alignments, and a multi-MSA strategy for high-quality protein monomer and complex structure prediction in CASP15. Proteins 2023;91:1684–703. https://doi.org/10.1002/prot.26585.
[4] Wallner B. AFsample: improving multimer prediction with AlphaFold using massive sampling. Bioinformatics 2023;39:btad573. https://doi.org/10.1093/bioinformatics/btad573.
[5] Peng Z, Wang W, Wei H, Li X, Yang J. Improved protein structure prediction with trRosettaX2, AlphaFold2, and optimized MSAs in CASP15. Proteins 2023;91:1704–11. https://doi.org/10.1002/prot.26570.
[6] Abramson J, Adler J, Dunger J, Evans R, Green T, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024;630:493–500. https://doi.org/10.1038/s41586-024-07487-W.
[7] Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)—round XV. Proteins 2023;91:1539–49. https://doi.org/10.1002/prot.26617.
[8] Alexander LT, Durairaj J, Kryshtafovych A, Abriata LA, Bayo Y, et al. Protein target highlights in CASP15: analysis of models by structure providers. Proteins 2023;91:1571–99. https://doi.org/10.1002/prot.26545.
[9] Roy RS, Liu J, Giri N, Guo Z, Cheng J. Combining pairwise structural similarity and deep learning interface contact prediction to estimate protein complex model accuracy in CASP15. Proteins 2023;91:1889–902. https://doi.org/10.1002/prot.26542.
[10] Edmunds NS, Alharbi SMA, Genc AG, Adiyaman R, McGuffin LJ. Estimation of model accuracy in CASP15 using the ModFOLDdock server. Proteins 2023;91:1871–8. https://doi.org/10.1002/prot.26532.
[11] Olechnovič K, Valančauskas L, Dapkunas J, Venclovas. Prediction of protein assemblies by structure sampling followed by interface-focused scoring. Proteins 2023;91:1724–33. https://doi.org/10.1002/prot.26569.
[12] Olechnovič K, Venclovas. VoroIF-GNN: Voronoi tessellation-derived protein-protein interface assessment using a graph neural network. Proteins 2023;91:1879–88. https://doi.org/10.1002/prot.26554.
[13] Liu J, Liu D, Zhang G. DeepUMQA3: a web server for accurate assessment of interface residue accuracy in protein complexes. Bioinformatics 2023;39:btad591. https://doi.org/10.1093/bioinformatics/btad591.
[14] Studer G, Tauriello G, Schwede T. Assessment of the assessment—all about complexes. Proteins 2023;91:1850–60. https://doi.org/10.1002/prot.26612.
[15] Zhao K, Zhao P, Wang S, Xia Y, Zhang G. FoldPAthreader: predicting protein folding pathway using a novel folding force field model derived from known protein universe. Genome Biol 2024;25:152. https://doi.org/10.1186/s13059-024-03291-x.
[16] McConkey BJ, Sobolev V, Edelman M. Discrimination of native protein structures using atom–atom contact scoring. Proc Natl Acad Sci 2003. https://doi.org/10.1073/pnas.0535768100.
[17] Goede A, Preissner R, Frömmel C. Voronoi cell: New method for allocation of space among atoms: elimination of avoidable errors in calculation of atomic volume and density. J Comput Chem 1997;18:1113–23. https://doi.org/10.1002/(SICI)1096-987X(19970715)18:9< 1113::AID-JCC1> 3.0.CO;2-U.
[18] Olechnovič K, Venclovas Č. VoroMQA: assessment of protein structure quality using interatomic contact areas. Proteins 2017;85:1131–45. https://doi.org/10.1002/prot.2527.
[19] Eisenberg D, McLachlan A. Solvation energy in protein folding and binding. Nature 1986;319:199–203. https://doi.org/10.1038/319199a0.
[20] VcG Satorras, Hoogeboom E, Welling M. E(n) Equivariant graph neural networks. Presente Proc 38th Int Conf Mach Learn Proc Mach Learn Res 2021:9323–32. ⟨https://proceedings.mlr.press/v139/satorras21a.html⟩.
[21] Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 2013;29:2722–8. https://doi.org/10.1093/bioinformatics/btt473.
[22] Basu S, Wallner B. DockQ: a quality measure for protein-protein docking models. PLoS One 2016;11:e0161879. https://doi.org/10.1371/journal.pone.0161879.
[23] Mirabello C, Wallner B. DockQ v2: improved automatic quality measure for protein multimers, nucleic acids, and small molecules. Bioinformatics 2024;40:btae586. https://doi.org/10.1093/bioinformatics/btae586.
[24] Olechnovič K, Kulberkytė E, Venclovas Č. CAD-score: a new contact area difference-based function for evaluation of protein structural models. Proteins 2013;81:149–62. https://doi.org/10.1002/prot.24172.
[25] Schrodinger, L.L.C., The PyMOL Molecular Graphics System, Version 2.6, ⟨www.pymol.org⟩.
[26] Pagès G, Charmettant B, Grudinin S. Protein model quality assessment using 3D oriented convolutional neural networks. Bioinformatics 2019;35:3313–9. https://doi.org/10.1093/bioinformatics/btz122.
[27] Hiranuma N, Park H, Baek M, Anishchenko I, Dauparas J, Baker D. Improved protein structure refinement guided by deep learning based accuracy estimation. Nat Commun 2021;12:1340. https://doi.org/10.1038/s41467-021-21511-x.
[28] Ballester PJ, Richards WG. Ultrafast shape recognition to search compound databases for similar molecular shapes. J Comput Chem 2007;28:1711–23. https://doi.org/10.1002/jcc.20681.
[29] Guo S, Liu J, Zhou X, Zhang G. DeepUMQA: ultrafast shape recognition-based protein model quality assessment using deep learning. Bioinformatics 2022;38:1895–903. https://doi.org/10.1093/bioinformatics/btac056.
[30] Liu J, Liu D, He G, Zhang G. Estimating protein complex model accuracy based on ultrafast shape recognition and deep learning in CASP15. Proteins 2023;91:1861–70. https://doi.org/10.1002/prot.26564.
[31] 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 Bioinf 2023;25. https://doi.org/10.1093/bib/bbad420.
[32] Shanker VR, Bruun TUJ, Hie BL, Kim PS. Unsupervised evolution of protein and antibody complexes with a structure-informed language model. Science 2024;385:46–53. https://doi.org/10.1126/science.adk8946.
[33] Lin Z, Akin H, Rao R, Hie B, Zhu Z, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv 2022;2022:500902. https://doi.org/10.1101/2022.07.20.500902.

[34] Hsu C., Verkuil R., Liu J., Lin Z., Hie B., Learning inverse folding from millions of predicted structures presente, Proc 39th Int Conf Mach Learn, Proc Mach Learn Res2022. ⟨https://proceedings.mlr.press/v162/hsu22a.html⟩. 16289468970.
[35] Lin Z, Akin H, Rao R, Hie B, Zhu Z, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379:1123–30. https://doi.org/10.1126/science.ade2574.
[36] Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzym 2011;487:545–74. https://doi.org/10.1016/B978-0-12-381270-4.00019-6.
[37] Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using Rosetta. Methods Enzym 2004;383:66–93. https://doi.org/10.1016/S0076-6879(04)83004-0.
[38] Uziela K, Shu N, Wallner B, Elofsson A. ProQ3: improved model quality assessments using Rosetta energy terms. Sci Rep 2016;6:33509. https://doi.org/10.1038/srep33509.
[39] He K., Zhang X., Ren S., Sun J. 2016. Identity mappings in deep residual networkscomputer vision–ECCV: 14th European Conference, Amsterdam, The Netherlands, October 11–14, Proceedings, Part IV 1420162016, Springer International Publishing. 201663064510.1007/978-3-319-46493-0_38.
[40] Basha SHS, Dubey SR, Pulabaigari V, Mukherjee S. Impact of fully connected layers on performance of convolutional neural networks for image classification. Neurocomputing 2020;378:112–9. https://doi.org/10.1016/j.neucom.2019.10.008.
[41] Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. Neurocomputing 2021;452:48–62. https://doi.org/10.1016/j.neucom.2021.03.091.
[42] Hendrycks D., Gimpel K. (2016) Gaussian Error Linear Units (GELUs). arXiv preprint arXiv:1606.08415. https://doi.org/10.48550/arXiv.1606.08415.
[43] Zhang Z. Improved Adam optimizer for deep neural networks. IEEE/ACM 26th Int Symp Qual Serv (IWQoS) Ieee 2018;2018:1–2. https://doi.org/10.1109/IWQoS.2018.8624183.
[44] Bagaria A, Jaravine V, Huang YJ, Montelione GT, Guntert P. Protein structure validation by generalized linear model root-mean-square deviation prediction. Protein Sci 2012;21:229–38. https://doi.org/10.1002/pro.2007.
[45] De Winter JC, Gosling SD, Potter J. Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: a tutorial using simulations and empirical data. Psychol Methods 2016;21:273. https://doi.org/10.1037/met0000079.
[46] Ali Abd Al-Hameed K. Spearman's correlation coefficient in statistical analysis. Int J Nonlinear Anal Appl 2022;13:3249–55. https://doi.org/10.22075/ijnaa.2022.6079.
[47] Ling C.X., Huang J., Zhang H., AUC: a better measure than accuracy in comparing learning algorithms, advances in artificial intelligence: 16th Conference of the Canadian Society for Computational Studies of Intelligence, AI 2003, Halifax, Canada, June 11-13, 2003, Proceedings 16. Springer Berlin Heidelberg, 2003: 329-341. https://doi.org/10.1007/3-540-44886-1_25.
[48] Margenau H. Van der waals forces. Rev Mod Phys 1939;11:1–35. https://doi.org/10.1103/RevModPhys.11.1.
[49] Dzyaloshinskii IE, Lifshitz EM, Pitaevskii LP. The general theory of van der Waals forces. Adv Phys 1961;10:165–209. https://doi.org/10.1080/00018736100101281.
[50] Krapp LF, Abriata LA, Cortés Rodriguez F, Dal Peraro M. PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. Nat Commun 2023;14:2175. https://doi.org/10.1038/s41467-023-37701-8.