Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Mini-review

Recent advances and challenges in protein complex model accuracy estimation

Fang Liang [1], Meng Sun [1], Lei Xie, Xuanfeng Zhao, Dong Liu, Kailong Zhao, Guijun Zhang *

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

ARTICLE INFO

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

ABSTRACT

Estimation of model accuracy plays a crucial role in protein structure prediction, aiming to evaluate the quality of predicted protein structure models accurately and objectively. This process is not only key to screening candidate models that are close to the real structure, but also provides guidance for further optimization of protein structures. With the significant advancements made by AlphaFold2 in monomer structure, the problem of single-domain protein structure prediction has been widely solved. Correspondingly, the importance of assessing the quality of single-domain protein models decreased, and the research focus has shifted to estimation of model accuracy of protein complexes. In this review, our goal is to provide a comprehensive overview of the reference and statistical metrics, as well as representative methods, and the current challenges within four distinct facets (Topology Global Score, Interface Total Score, Interface Residue-Wise Score, and Tertiary Residue-Wise Score) in the field of complex EMA.

1. Introduction

Proteins serve as the foundation for sustaining life activities and cellular functions, playing a crucial role in the majority of biological processes [1]. Understanding the tertiary structure of the protein is decisive for comprehending its function and its interactions with other molecules. With the development of artificial intelligence, protein structure prediction has made great progress. Estimation of model accuracy (EMA) is an important part of protein structure prediction. The efficient EMA method may identify errors in experimental structures and reflect the reliability of prediction models, thereby guiding model refinement and promoting the development of innovative drug and vaccine design based on structure. EMA methods constitute a crucial category of the Critical Assessment of Protein Structure Prediction (CASP) experiments [2–4]. EMA was first introduced in CASP7 (2006), emphasizing the significance of EMA in protein tertiary structure prediction [5]. Since 2012, another important worldwide competition Continuous Automated Model EvaluatiOn (CAMEO) [6] has introduced a weekly online automated blind evaluations of protein structure prediction servers and EMA servers, which complements the biennial CASP experiments and accelerating the advancement of EMA methods.

The EMA methods can be roughly divided into consensus methods, quasi-single model methods and single model methods [7]. Based on the assumption that the correct structure information is contained in the repeated structural patterns of the model pool, consensus methods extract consensus information from protein structure models through clustering. Representative methods include the MULTICOM series [8,9], the MUfoldQA series [10,11], the ModFOLDclust series [12,13], clustQ [14], Pcons [15], APOLLO [16]. From the CASP results, it appears that consensus methods perform better than single model methods in most protein targets [7]. Quasi-single model methods take a single protein model as input and compare it for structural similarity with a set of internally generated models. Representative methods include the ModFOLD series [17–19], QMEANDisco [20]. The single model method extracts the sequence, geometric structure, and physical and chemical features of a single protein model, and inputs it into the neural network to predict the quality of local residues or global topology. With the rapid development of deep learning technology in the field of protein structure prediction, the performance of single model methods has gradually equaled or even surpassed consensus methods, becoming a research hotspot for EMA. The representative methods mainly include the Voro series [21–23], DeepUMQA series [24,25], GraphQA [26], AlphaFold2 [27], DeepAccNet series [28], QDistance [29], Qdeep [30], Ornate [31], AngularQA [32], ProQ series [33–40], 3DCNN [41], QAcon [42],

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

[1] The first two authors should be regarded as Joint First Authors.

https://doi.org/10.1016/j.csbj.2024.04.049
Received 27 January 2024; Received in revised form 18 April 2024; Accepted 18 April 2024
Available online 21 April 2024

SVMQA [43], DeepQA [44], QMEAN [45], ProSA [46,47]. These monomer EMA methods are of great significance to the EMA field and lay the foundation for complex EMA. More detailed descriptions of these methods can be found in the reviews [48–50].

With AlphaFold2 (AF2) has made significant breakthroughs in predicting protein tertiary structures and self-assessment, researchers have shifted their attention from monomers to protein complexes [27,51,52]. Currently, the accuracy of protein complex modeling is much lower than that of monomer modeling [7]. Complex EMA methods are critical to improve their prediction accuracy. Therefore, the development of EMA methods for complexes has become particularly important. The CASP15 EMA methods have shifted the focus on interfaces to highlight the importance of protein-protein interactions in understanding the function and stability of quaternary structure [7]. In addition, the Critical Assessment of PRediction of Interactions (CAPRI) [53] aims to evaluate the ability of protein docking methods to predict protein-protein interactions. CAPRI has established a close relationship with CASP and successfully held the fifth joint CASP-CAPRI experiment [54–58]. Complex EMA involves the comprehensive evaluation using various metrics. Specifically, reference metrics refers to the various metrics and scores used to assess the accuracy of a structural model whereas statistical metrics are used to evaluate the capabilities of different methods for predicting such metrics and scores. The complex assessment in CASP15 includes three distinct tracks: SCORE, QSCORE, and Local [7]. For the SCORE track, reference metrics Oligo-GDTTS [7,59] and TM-score [60] are used to evaluate the overall topology of complexes, with the state-of-the-art methods being MULTICOM_qa [61], developed by Jianlin Cheng's research group. The QSCORE track uses reference metrics DockQ-wave [7,62] and QS-score [63] to assess the quality of the interface, with the top-ranking methods including the ModFOLDdock series [64] developed by the research group led by McGuffin. For the Local track, reference metrics lDDT [65], CAD-score [66] and the newly proposed PatchDockQ [62] and PatchQS [63] are employed to assess the accuracy of interface residues in each model, where the top-ranking research groups include the DeepUMQA3 [67], developed by Guijun Zhang's research group. Furthermore, the performance of EMA methods was estimated according to several statistical metrics, such as Pearson [68], Spearman [68], and AUC ROC [3,69]. These statistical metrics were then weighted and transformed into a Z-score for the purpose of scoring and ranking [7].

2. Methods of complex model accuracy estimation

2.1. Overview

With significant advances in monomer structure prediction, complex model accuracy estimation has become a research hotspot. In order to comprehensively evaluate the complex model accuracy, various reference and statistical metrics have been proposed. Based on the composition and functions of the quaternary structure, we classify and describe them through four facets: Topology Global Score (TGS), Interface Total Score (ITS), Interface Residue-Wise Score (IRWS), and Tertiary Residue-Wise Score (TRWS) (Fig. 1). TGS assesses the accuracy of the topological structure, ITS evaluates the precision of inter-molecular interactions within the complex, IRWS focuses on the accuracy of individual residue at the interface, and TRWS highlights the quality of per-residue in tertiary structure. Representative methods in each facet are listed in

Table 1
Brief description of reference metrics, statistical metrics, representative methods and CASP15 performance of EMA in four facets. Based on the data from the CASP15 official website (https://predictioncenter.org/casp15), the sum of Z-scores for the CASP15 representative methods is calculated by all metrics, which can be used to rank the prediction accuracy of methods.

Facets	Reference metrics	Statistical metrics	Representative methods	CASP15 performance (Z-score)
Topology Global Score (TGS)	TM-score Oligo-GDTTS	Pearson Spearman ROC AUC Top1 loss	MULTICOM_qa ModFOLDdock series VoroIF-jury GraphGPSM	7.555 7.095 6.487 3.003
Interface Total Score (ITS)	QS-score DockQ DockQ-wave	Pearson Spearman ROC AUC Top1 loss	ModFOLDdock series VoroIF-GNN DeepUMQA3	7.286 4.043 3.789
Interface Residue-Wise Score (IRWS)	lDDT CAD-score PatchQS PatchDockQ	Pearson Spearman ROC AUC	DeepUMQA3 ModFOLDdock series VoroIF-GNN VoroIF-jury GraphCPLMQA	13.950 8.898 6.971 6.948 /
Tertiary Residue-Wise Score (TRWS)	lDDT	Pearson ASE ROC AUC PR AUC	AlphaFold2 ColabFold DeepUMQA1,2	/ / /

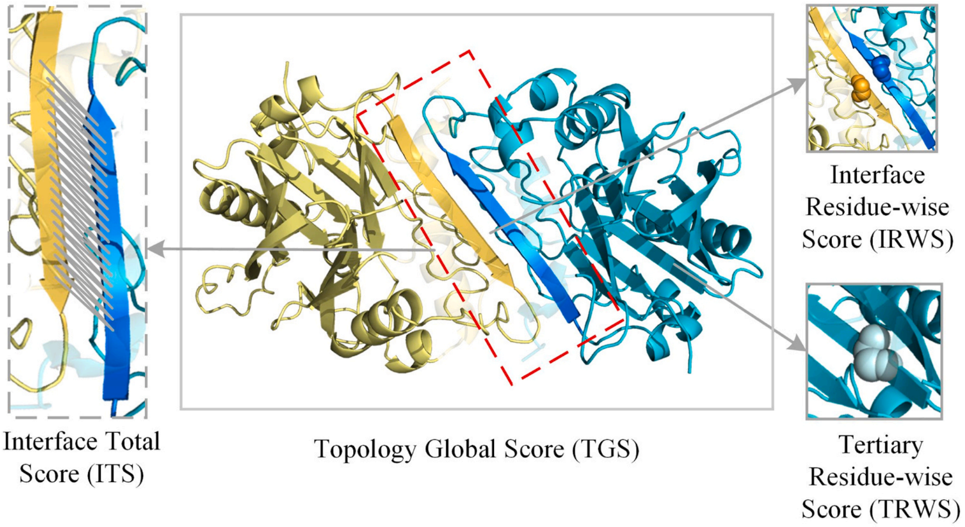
Interface Total Score (ITS)
Topology Global Score (TGS)
Interface Residue-wise Score (IRWS)
Tertiary Residue-wise Score (TRWS)

Fig. 1. Schematic diagram of the four quality assessment facets defined as Topology Global Score (TGS), Interface Total Score (ITS), Interface Residue-Wise Score (IRWS), and Tertiary Residue-Wise Score (TRWS) of the complex.

Table 1. The definition of common metrics is represented in supplementary material. For these four facets, this section introduces the reference metrics, statistic metrics and their characteristics. Then we summarize the main advantages of the representative methods in the past two years, and discuss the key challenges and further development directions in the field.

2.2. Topology global score (TGS)

Assessing the topological similarity between prediction models and the native structure of proteins is crucial for ensuring the biological reliability and practicality of these models. The TGS facet assessment of complexes usually employed the reference metrics Oligomer Global Distance Test Total Score (Oligo-GDTTS) [7] and the Template Modeling Score (TM-score) [60]. Oligo-GDTTS is conceptually similar to GDT [70] for complex, which calculates the maximum number of atoms in the modeled structure corresponding to the experimental structure within a specified threshold (i.e.1, 2, 4, 8 Å) by superimposing the predicted structure with the experimental reference structure. It does not penalize residues over the threshold range [59,60]. The value range of Oligo-GDTTS is usually between 0 and 1. The closer the score is to 1, the more similar the prediction model is to the reference structure. TM-score is a widely used reference metric for assessing the quality of protein structure models, which eliminates the dependence on protein length in the previous evaluation metrics by using protein size-dependent values [60]. It is employed to evaluate all residue pairs in alignment, instead of setting a specific distance cutoff and computing error scores below the cutoff. As a result, TM-score focuses on global topology features and may ignore local structural details [71]. Two monomer protein structures with a TM-score> 0.5 are considered to have the same topology [71]. For complex models, they are considered acceptable quality if the TM-score is above 0.7 and high quality if the TM-score is above 0.8 [72]. When the reference structure is unknown, the predicted TM-score (pTM) [27] derived from AF2 assumes the existence of a distribution of probable structures and uses the pairwise error matrix to find the expected value of the TM-score for the predicted structure. Considering the different advantages of Oligo-GDTTS and TM-score in EMA, they are often used in combination in CASP to meet the need for a more comprehensive and accurate assessment of protein models. To further objectively and fairly analyze the accuracy and reliability of the assessment, statistical metrics Pearson, Spearman, the area under the Receiver Operating Characteristic curve (ROC AUC), and Top1 loss are employed to ensure overall performance on all targets. Particularly, Top1 loss refers to the difference between the selected top model and the true best model for the target protein and focuses on ranking the optimal model [73]. These statistical metrics collectively reflect the comprehensive performance of evaluation methods.

In CASP15, we have witnessed recent advances in the complex EMA methods. Consensus methods are in a domination state in the performance assessment of TGS facet, which extract highly consistent structural information from a multitude of different models and apply a combination of evaluation strategies that are functionally complementary to each other. The leading consensus methods include MULTICOM_qa [61] of Jianlin Cheng's research group, VoroIF-jury [74] of Venclovas's research group, ModFOLDdock [64] of McGuffin's research group, and so on [29]. MULTICOM_qa combines pairwise similarity scores (PSS) and interface contact probability scores (ICPS) to predict the global accuracy of complex structure [61]. The average Pearson of the MULTICOM_qa method is 0.683, which is 14.6 % higher than that of using an assembly consensus baseline predictor known as "AC" (only structure comparison score), and achieves an average Top1 loss of 0.152 on 40 targets, which is 14.1 % lower than that of "AC" [61,75]. This result shows that MULTICOM_qa is able to effectively assess the overall quality of the model, reasonably select the best model, and perform best in the CASP15 SCORE track [61]. VoroIF-jury specializes in protein assembly modeling and uses a scoring strategy that focuses on interface accuracy, much of which comes from the VoroMQA interface energy ranking models predicted by various modeling methods [74]. In addition, ModFOLDdock aims to enhance the correlation between predicted and reference scores, placing second in the SCORE track at CASP15 [64]. Despite the excellent performance of consensus methods, there are obvious limitations. Firstly, consensus methods require high computational costs to obtain consensus information on a large number of structures by comparing several different models, especially for large proteins. Secondly, the extensive integration of AF2 in most of the current algorithms leads to significant similarities between the pool of candidate models and a strong dependence on the accuracy of the AF2 models. Single model methods employed by participants in CASP15 appeared less constrained by these potential influences. The running times of representative consensus methods and single model methods are illustrated in Fig. S1. Obviously, the computational costs of single model methods are much lower than consensus methods.

Recently, with the advancements of deep learning, single model methods have become increasingly important. Single model methods have gradually caught up with or surpassed the consensus method [7]. Single model methods assess models without the need for using structural information from other server models. The single model method GraphGPSM [77] (group name: GuijunLab-Threader) participated in CASP15, which was designed to guide protein structure modeling and selecting. This method based on equivariant graph neural network (EGNN), achieves information interaction between graph nodes and edges through a message passing mechanism [77]. CASP15 blind test results demonstrated a strong correlation between the predicted and actual TM-score of the models. GraphGPSM predicted 35 targets in CASP15, and the mean absolute error (MAE) of the predictions based on the TM-score in relation to the native structure was 0.126, which was the smallest average bias among top-ranked servers [78]. Since GraphGPSM primarily relies on TM-score, many improvements are still needed to capture the interface relationships of complexes and score them accurately. GraphGPSM is at the leading status among all the single model methods entered in CASP15.

2.3. Interface total score (ITS)

Protein-protein interfaces refer to the contact surface or binding region between two or more proteins. As a key part of protein-protein interactions, the properties of the protein-protein interface affect many important processes that regulate cellular functions, such as signal transduction, metabolism, immune responses, and so on [79]. Therefore, specialized metrics must be designed to measure the protein-protein interface prediction quality of the prediction models (illustrated in Fig. 1, ITS). As two important reference metrics to evaluate ITS, DockQ [62] and QS-score [63] have been widely used in CASP experiments and CAPRI rounds [54,57,58].

DockQ primarily focuses on binary interactions, which evaluates the accuracy of protein docking models by integrating scores of Fnat, LRMS, and iRMS [62,80,81]. Fnat measures the proportion of contact residues in the predicted complex interface that match those in the reference complex interface. The interface is defined as the region within 5 Å of any pair of heavy atoms in two interacting molecules. LRMS is calculated as the Root Mean Square Deviation for the model's ligand, after superimposing the receptors of the predicted and reference complexes. For iRMS, the atomic contact cutoff for the receptor-ligand interface in the target is redefined to 10 Å [62]. The quality of models with a DockQ below 0.23 is considered as "incorrect", between 0.23 and 0.49 as "acceptable", between 0.49 and 0.8 as "medium", and above 0.8 as "high" [62]. The predicted DockQ score (pDockQ) [82] uses a combination of the average predicted lDDT (plDDT) of interface residues and the logarithm of the number of interface contacts to fit the observed DockQ. pDockQ can accurately assess protein docking quality when the native structure is unknown. Its improved version, pDockQ2 [83], uses the predicted alignment error (PAE) of all interfaces to quantify the quality of each interface of a multimer and accurately distinguish those large, high-confidence but incorrect interfaces. In addition, the interface pTM (ipTM) [84] derived from AlphaFold-Multimer (AFM) is an interface version of pTM, which greatly advances the high-confidence prediction of protein-protein interactions. For higher-order complexes containing more than two interfaces, it is clearly not sound to average each interface to derive its total score. In this case, DockQ-wave was introduced to provide a more accurate assessment of complex interfaces, which weights the DockQ scores based on the number of distinct interface residues in the native structure [7,62]. DockQ-wave splits the complex into dimers for one-by-one analysis, and in some cases may have missed some interface contact residues in the overall structure. To alleviate this challenge, the QS-score was used as an auxiliary reference metric to evaluate the interface accuracy by overall assessment of the shared contacts of the high-order complex [63]. The evaluation employs different cutoff values for this metric, classifying prediction models into four quality categories: Incorrect (QS-score < 0.1), Low (QS-score: 0.1–0.3), Medium (QS-score: 0.3–0.7), and High (QS-score > 0.7) [63]. Considering the complementary advantages between DockQ and QS-score, CASP15 integrated the two metrics to evaluate ITS. In addition, statistical metrics similar to TGS, such as Pearson, Spearman, ROC AUC, and Top1 loss, were used to measure and rank the predictive performance of the different methods for all targets. Generally, with the development of protein structure prediction and deep learning technology, great progress has been made in ITS accuracy estimation methods. In recent years, ModFOLDock series of McGuffin research group [64], VoroIF-GNN of Venclovas research group [76] and some other representative methods [61,85,86] have emerged.

ModFOLDdock brings together a range of single model, clustering, and deep learning methods to form consensus assessments, which was optimized for positive linear correlations with observed scores [64]. For assessing different facets of model quality, two variants of ModFOLDdock, namely ModFOLDdockR and ModFOLDdockS, have been developed. ModFOLDdockR adds the VoroMQA method [23] based on the standard version, improving the ability to select top1-ranked models. In CASP15, ModFOLDdockR outperformed other methods except “AC” on the Pearson of ITS, which highlights the ability of ModFOLDdockR to accurately assess the quality of complex protein interactions. Although ModFOLDdockR shows good performance on most targets, it relies heavily on the model pool of CASP scenario [64]. To overcome this problem, the quasi-single model method ModFOLDdockS constructs a reference model pool using a set of protein structures generated by their own developed MultiFOLD method [87]. ModFOLDdockS combines the convenience of a single model as input with the evaluation ability of consensus methods based on generating ensemble information, providing a more robust and accurate score for protein interface quality evaluation. While the ModFOLDdock series exhibited leading performance in the ITS during CASP15, it is noteworthy that it has yet to surpass the consensus baseline method “AC” [7]. This demonstrates that there is still considerable room for improvement in the quality assessment of ITS. Moreover, with the ongoing advancements in deep learning technology, single model methods may eventually surpass the current baseline method.

The single model method VoroIF-GNN derives interface contacts from the Voronoi tessellation of atomic balls to construct the graph and predict the accuracy of each contact using an attention-based GNN [76]. It shows better performance compared to single model methods of other groups in ITS of CASP15, and the performance of the single model method DeepUMQA3 [67] is comparable to VoroIF-GNN. In particular, VoroIF-GNN and DeepUMQA3 outperform consensus methods on most nanobody-antigens and proteins with structural flexibility on Pearson [7]. The reason is that the rapid evolution of viral protein sequences may hinder multiple sequence alignment (MSA) [88], resulting in less accurate structural modeling. This highlights the great potential of single model methods, which require only a single protein complex structure as input and do not directly use additional information (e.g., MSA) to accurately evaluate and select antibody-antigen complex proteins. It is worth noting that MULTICOM_qa achieved the best performance on TGS of CASP15, while ranking only 16th on ITS [75]. This suggests that TGS and ITS characterize complexes from different perspectives, and it may not be possible to comprehensively assess the quality of protein models using only one method. Based on the current advances of EMA and the results of the CASP15 blind test, it can be concluded that in terms of accurately evaluating inter-molecular interactions, current complex EMA methods face difficulties in surpassing the baseline “AC”; the performance of consensus methods is better than that of single model methods in most proteins; single model methods demonstrate the potential to surpass consensus methods on antibody-antigens and proteins with structural flexibility.

2.4. *Interface residue-wise score (IRWS)*

The interface residues of protein-protein are critical for maintaining the structural stability of protein complexes. Mutations in interface residues may affect the interactions of protein complexes, leading to changes in their functions. In order to accurately evaluate the quality of local residues on the protein interface, the local Distance Difference Test (lDDT) [65] and the contact area difference-based score (CAD-score) [66] are used in CASP or CAMEO, which are based on contacts to assess the difference of the relative positions of neighboring atoms between the prediction model and the native structure.

lDDT can accurately assess the local geometry at protein binding sites by calculating the local distance difference of the atomic pairs between the model and the reference structure at specific distance thresholds (e.g., 0.5 Å, 1 Å, 2 Å, 4 Å) [65]. CAD-score utilizes the van der Waals radii to consider the size of the atoms, and quantifies the contact from a physical viewpoint by calculating the difference in contact area of the residues between the model and the reference structure [66]. The value range of lDDT and CAD-score is usually between 0 and 1. The closer the score is to 1, the closer the prediction model is to the native structure. lDDT and CAD-score, originally designed for assessing the local predictive accuracy of tertiary structure, do not explicitly penalize additional interface contacts such as residues that should be at the surface but are mistakenly modeled on the interface and their calculations may be primarily influenced by intrachain contacts [7,89]. To alleviate these problems, PatchDockQ and PatchQS introduced in CASP15, generate two local patches for each interface residue [7]. These patches pair with their target structure counterparts and form dimers to calculate DockQ and QS-score, providing a more targeted assessment of interchain contacts. Considering the complementary strengths of lDDT, CAD-score, PatchDockQ, and PatchQS, CASP15 combines these metrics to evaluate IRWS. Unlike TGS and ITS, which are often used for model selection, IRWS focuses more on guiding protein structure refinement [28]. Therefore, CASP15 does not employ the statistical metric Top1 loss in IRWS, but only Pearson, Spearman, and ROC AUC to measure the reliability of predictions regarding interface residues. Representative methods include the DeepUMQA series [24,25,67] of Guijun Zhang’s research group, the ModFOLDdock series [64] of McGuffin’s research group and the VoroIF-GNN [76] and VoroIF-jury [74] of Venclovas’ research group, which have conducted significant work in this facet.

The single-model method DeepUMQA3 (group name: GuijunLab-RocketX) [67] introduces residue-level ultrafast shape recognition (USR) to capture the relationship between residues and the overall protein topology, and then combines 1D features, 2D features, and voxelized embeddings to evaluate the quality of interface residues. These features are fed into a residual neural network that combines triangle updating and axial attention for predicting each residue’s lDDT score [67]. The method ranked 1st in the IRWS (Local track) of CASP15 and showed best performance on all four reference metrics lDDT, CAD-score, PatchDockQ and PatchQS. Notably, DeepUMQA3 achieved the highest Pearson of lDDT on three of the five nanobody-antigens and all three antibody-antigens, which were hard to predict in the CASP experiments [78]. Considering that a connection exists between sequence, structure and quality, a single model method GraphCPLMQA [50] based on DeepUMQA series is further proposed, which uses embeddings from the protein language model ESM [90] and a deep graph-coupled network to assess residue-level protein model quality. For both five nanobody-antigens and three antibody-antigens, the GraphCPLMQA method enhanced the average Pearson based on lDDT by 36.5 % and 8.6 %, respectively, compared to DeepUMQA3 [50,67] . In addition, the ModFOLDdock series demonstrate strong performance in accurately identifying interface residues and assessing interface residues of antibody-antigens [64]. Other representative methods that demonstrate comparable performance to ModFOLDdock series are the single model method VoroIF-GNN [76] and the consensus method VoroIF-jury [74].

Accurately assessing the quality of interface residues is a critical task, requiring not only the determination of whether these residues are actually part of the interface but also the evaluation of the modeling accuracy of these interface residues. The progress of EMA methods in IRWS and the results of CASP15 show that while single model methods are well beyond consensus methods [7], obvious disparities in predictive performance exist across various targets, which suggests that there is still room for improvement in single model methods. Future EMA methods should focus on accurately identifying the true residues at the interface to improve prediction accuracy and better guide complex modeling.

2.5. *Tertiary residue-wise score (TRWS)*

Although AF2 has made breakthroughs in tertiary structure prediction, there are still challenges in multi-domain protein assembly and the structure modeling of proteins with multiple conformations. Thereby, as an indispensable part of tertiary structure prediction, the self-assessment of tertiary structure remains a research topic worthy of our attention, which identifies high-confidence regions of prediction models, providing guidance for applications of model refinement [91], molecular replacement [92], and template identification [93]. In self-assessment of tertiary structure, a widely used reference metric lDDT is used in TRWS, which accurately reflects the precision of the atomic environment around each residue in the model [65]. In addition, the statistical metrics Pearson, ROC AUC, the area under the precision-recall curve (PR AUC), and Accuracy Self Estimate (ASE) were used in combination to analyze the performance of different methods on TRWS in CASP15. The definition of ASE is in the Supplementary Text S8. Based on the official results of tertiary structure self-assessment of CASP15, the current methods perform the same as or better than the consensus method [7], such as ColabFold [94], AF2 [27], and DeepUMQA2 [25].

ColabFold of Steinegger's research group [94] accelerates prediction of protein structures through combining the fast homology search of MMseqs2 [95] with AF2, which ranks 1st in tertiary structure self-assessment on the sum of the four statistical metrics (Pearson's r, ASE, ROC & PR AUC) in CASP15. The other top-performing methods also include FoldEver, AF2, and MUFold [96], providing accurate per-residue confidence estimates at performance comparable to consensus method [7]. It is worth noting that the performance differences between these methods are minimal. This result may be attributed to the fact that most of the top-performing groups integrated AF2 into their methods to some degree, where AF2 not only predict highly accurate models but also provide reliable per-residue confidence estimates [7]. However, these AF2-dependent methods face challenges in ranking and selecting the best model among many similar poor models of hard targets. Therefore, it may be necessary to consider developing an evaluation method independent of modeling techniques as complementary method to address the challenge of ranking and selecting the best model.

Traditional single model methods are independent of modeling techniques and aim to provide self-assessment scores for tertiary structure based solely on the basic information of a single model. Early prominent single model methods mainly rely on knowledge-based statistical potentials. Representative methods such as ProSA [46,47] build mean force potentials on statistics of pairwise interaction distances to identify misfolded structures. Starting with ProQ [38], methods aimed at predicting the quality of protein models were developed. ProQ incorporates structural features, such as atom and residue contacts, solvent accessibility and predicted secondary structure consistency [38]. ProQres [39] expanded upon ProQ to estimate the quality of each residue. ProQ3D [33] uses the same input as ProQ3 [40] to train a multi layer perceptron model, which integrates Rosetta energy terms, including all-atom energy function and centroid energy function. ProQ4 [35] uses only coarse structural features and a multiple sequence alignment to improve estimation accuracy. Similar to ProQ, QMEAN [45] combines distance-dependent pairwise potential, solvation potential, torsion angle potential, secondary structure and solvent accessibility agreement to describe the major geometric features of protein structures. QMEANDisco [20] utilizes distance distribution of homologous model structures, and employs feedforward neural networks to weight the multi-template consensus-based distance constraint (DisCo) scores and QMEAN scores to obtain the final score. To further describe and analyze the physical properties of protein structures, VoroMQA [23] introduces statistical potentials based on atom-atom contact areas. Its improved version, VoroMQA-dark [97] uses three layers of residue neighborhood descriptors to train a feed-forward neural network to predict the pre-residue CAD-score. These methods are widely used in the EMA field and provide important references for subsequent EMA methods.

The single model methods DeepUMQA series [24,25,67] are independent of AF2. DeepUMQA utilizes USR features to characterize protein topology and employs a residual convolutional network to predict residue-wise model quality score. Based on DeepUMQA, DeepUMQA2 [24,25] further integrates protein sequence co-evolution information and protein template structural features, which are fed into an improved network with triangular multiplication update and axial attention mechanism to evaluate residue quality. In CASP15, "GuijunLab-Threader" performed structure modeling by replacing the template recognition component HHsearch [98] of AF2 with the in-house template search method PAthreader [99], and then used DeepUMQA2 to evaluate the model quality at the residue level. "GuijunLab-Threader" ranks 3rd overall out of 101 groups in the median ASE Z-score for the tertiary model self-assessment and ranks 1st among the server groups [100]. The result shows that DeepUMQA2 can accurately evaluate the residue-wise structure quality, surpassing AF2 and some methods derived from it. Moreover, "GuijunLab-Threader" ranks 6th out of 88 groups in selecting the best model [100]. It is worth noting that the success rate of all groups is less than 35% in selecting the best model, and only about 2/3 of them exceed the results of random selection (20% success rate) [7]. The main reason may be that the five models submitted by their respective groups are similar, which makes it difficult to select the best model. Further, this also suggests that ranking and selecting models by averages of residue-by-residue quality score may not be the optimal choice. In summary, great progress has been made in EMA for tertiary structure, but there are still challenges in selecting models. The development of EMA methods that are independent of modeling techniques is necessary for applications in downstream tasks, and may also be a promising way to verify the correctness of experimental structures in the future.

3. Challenges of complex model accuracy estimation

The widespread application of AF2 and AFM in the field of protein modeling has advanced the development of EMA technology, where consensus assessment methods perform well on most proteins [7]. With the development of deep learning, single model assessment methods have also made great progress. Nevertheless, current EMA techniques still face significant challenges in evaluating some complexes, such as antibody-antigens, flexible proteins, and large assemblies.

3.1. Antibody-antigen complex model accuracy estimation

EMA is crucial to guide the modeling and screen models of immunity complexes, which helps to accelerate antibody-based drugs development. In recent years, the evaluation of antibody-antigen complexes has attracted increasing attention. The results from CASP15 demonstrate significant challenges in the modeling of eight immunity complexes targets, including five nanobody-antigen and three antibody-antigen complexes [7]. This may be attributed to the fact that sequence mutations in the variable region (V region) of antibodies increase their diversity, and the highly variable complementarity determining region (CDR) loops enable antibodies to bind antigens with high specificity [74,101,102]. As show in Fig. 2 for H1140-H1144, multiple nanobodies bind to the same antigen target at different epitopes [98]. The "one-to-many" interaction mechanism brings great challenges to EMA technology of complex. Specifically, the average DockQ of prediction models of nanobody-antigen targets H1140-H1144 were 0.036, 0.070, 0.018, 0.430 and 0.079, respectively. Most poor models might directly lead to the failure of consensus assessment methods to accurately select models. In these targets, single model methods VoroIF-GNN and DeepUMQA3 successfully selected the best model for four and three out of the five nanobody-antigens, respectively, while "AC" only succeeded on H1143 (QS-score loss < 0.1) which the high-quality template exists [67,76]. With the progress of self-supervised learning in antibody modeling [101,103], we think that single model methods combined with pre-trained language models on natural antibody sequences (e.g., EATLM [104]) may be promising for accurately evaluating the quality of immunity complexes and selecting the best model.

3.2. Complex flexible residues model accuracy estimation

The biological function of a protein is determined not only by its static structure but also by the flexibility and dynamic properties of the complex. EMA of flexible complexes is critical to understanding the activity of proteins and their regulation mechanisms, which can help advance the development of biomedical research [105]. Most EMA methods are generally designed for static structures, which present a significant challenge to the evaluation of complexes with structural flexibility. Taking T1121o of CASP15 as an example (Fig. 3), which is a DNA-binding protein complex from Pseudomonas aeruginosa. During the process of binding to DNA, its binding pocket and active site appear to be exposed as the crossed DUF3322 domain moves [106], suggesting that this target has large structural flexibility. According to the blind test results of CASP15, the Pearson of T1121o for all groups were below 0.5 on QS-score, indicating that there is still large room for improvement in EMA technology for flexible complex. Moreover, the Pearson of T1121o for "AC" was 0.153 on the QS-score, while the single model method VoroIF-GNN achieved 0.402 [7,76]. The results show that the single model method outperforms the consensus method in the flexible complex, which may be due to the fact that the single model method captures the interaction mechanism of flexible complexes by incorporating physical and chemical features. To alleviate this challenge, we believe that single model methods may help improve evaluation performance by comprehensively considering the dynamic and physicochemical properties of flexible complex.

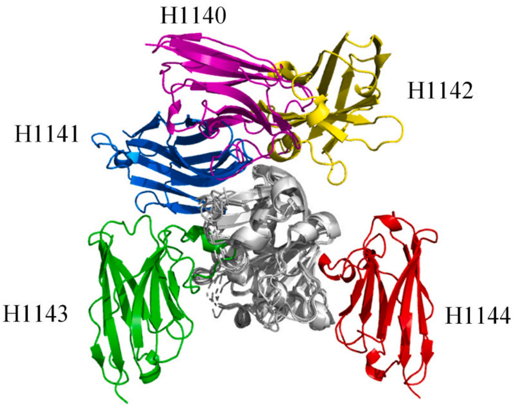

Fig. 2. Complex structure of five nanobody-antigen targets H1140-H1144 in CASP15, where the antigen is shown in gray and the five antibodies are shown in different colors. (All data in the figure are download from the official CASP15 website: https://predictioncenter.org/casp15).

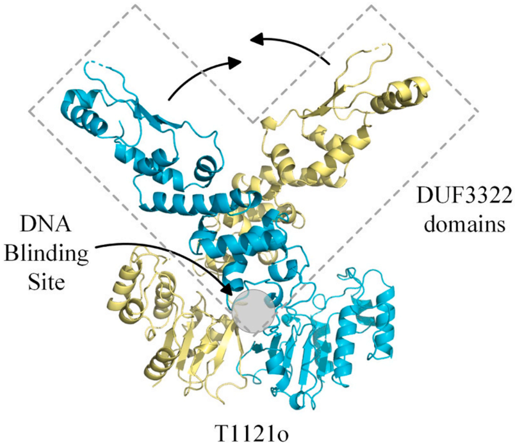

Fig. 3. Native structure of T1121o which represents an inactive conformation of the DNA-binding protein complex from Pseudomonas aeruginosa. The crossed DUF3322 domain is labeled with dashed boxes. The DNA-binding site is indicated with a gray circle (T1121o model cited from literature [106]).

3.3. Large assemblies model accuracy estimation

The evaluation of large assemblies presents significant additional challenges due to modeling difficulties and the consumption of enormous computational resources [78]. In CASP15, most consensus assessment methods performed well on large assemblies, which may be attributed to the fact that the model pool contains high-quality structures. However, these methods require "chain mapping" during the structure alignment process, which is a process of factorial complexity that greatly increases computational costs [7,61]. Taking H1111 (8460 residues) as an example (Fig. 4), it took more than three days to evaluate the model by the structurally aligned consensus method on a single CPU, resulting in only about half of the groups submitting evaluation scores. Similarly, while single model methods perform comparably to consensus methods when evaluating large assemblies with more than 3000 residues, they also face challenges and often exceed the memory limitations of a single GPU. In order to overcome computational costs and hardware resource limitations, we think that single model methods can adopt a "divide and conquer" strategy to rationally divide large models into smaller assembly units based on biological function for independent assessment. Furthermore, incorporating experimental data into the EMA technology may provide a more accurate assessment, which is essential for a comprehensive understanding of the structure and function of the large assemblies.

4. Conclusion

Estimation of model accuracy of protein complexes plays a crucial role in the structural biology. In this review, we retrospect the current progress in the field of complex EMA from four facets: Topology Global

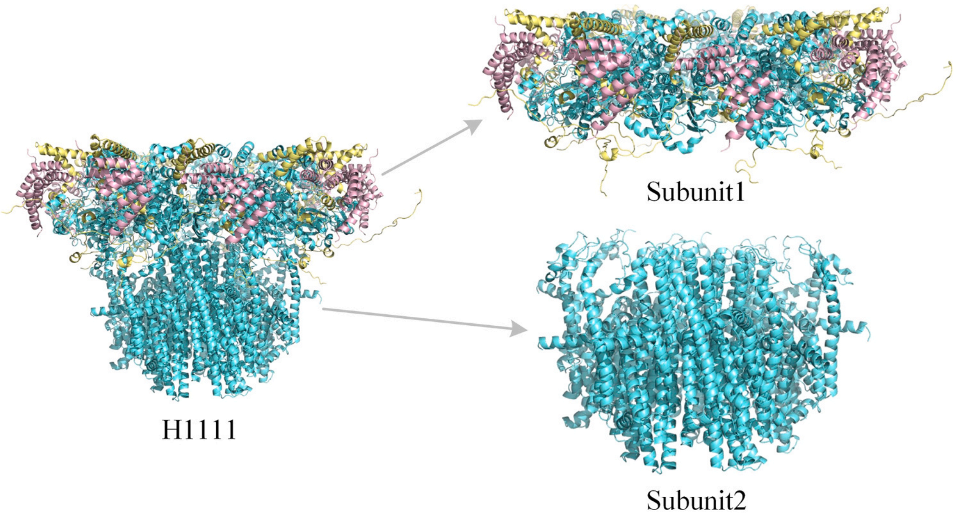

Fig. 4. Structure of the large assembly H1111 model (8460 residues) down from CASP15 official website, which is divided into two subunits through the "divide and conquer" strategy.

Score, Interface Total Score, Interface Residue-Wise Score, and Tertiary Residue-Wise Score. Further, we also discuss the challenges and provide insights that may address these challenges.

Based on the analysis of the CASP15 blind test results, we can conclude that the consensus methods perform well on most protein targets, but their performance significantly decreases on targets with a poor model pool [89]. In contrast, single model methods demonstrate advantages in accuracy assessment and selecting models on challenging protein targets, such as antibody-antigen structures and complexes with structural flexibility. With the development of artificial intelligence technology, we believe that the single model methods have the potential to surpass consensus methods by using deep learning to capture the mapping relationship between sequence and structure. This may also be the future development trend in the complex EMA field.

With the continuous advancements of structural modeling as well as EMA methods, many reference metrics have been developed and incorporated into evaluation systems to evaluate complex quality from different facets, which makes it difficult to rank the models or use them as loss functions in deep learning algorithms. It may be necessary to design a single robust reference metric covering different facets of the structural properties of the complex for EMA. Moreover, protein structures resolved by experimental techniques, such as cryo-EM, are not necessarily the true structure in the organism, especially for proteins with flexibility and multiple conformations. Notably, the structures deposited in the PDB may contain experiment errors, which may be identified by state-of-the-art EMA methods. Therefore, by integrating experimental techniques, the EMA computational methods can provide a more comprehensive understanding of protein structure and function. All in all, EMA is crucial for the development and benchmarking of protein structure prediction methods, we believe that this technology will have a broad, far-reaching, and long-lasting impact on the structural biology community, ultimately becoming an effective means of driving transformative technology in protein complex modeling.

CRediT authorship contribution statement

Guijun Zhang: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Formal analysis, Data curation, Conceptualization. **Fang Liang:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Lei Xie:** Writing – review & editing, Visualization, Methodology, Data curation. **Meng Sun:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Dong Liu:** Writing – review & editing, Visualization, Methodology, Data curation. **Xuanfeng Zhao:** Writing – review & editing, Visualization, Methodology, Data curation. **Kailong Zhao:** Writing – review & editing, Visualization, Methodology, Data curation.

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.

Acknowledgment

This study is supported by the National Science and Technology Major Project (2022ZD0115103), the National Nature Science Foundation of China (62173304).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.04.049.

References

[1] Matthews BW. Protein science best papers for 2020. Protein Sci: a Publ Protein Soc 2021;30(4):713–5. https://doi.org/10.1002/pro.4051.

[2] Kryshtafovych A, Fidelis K. Protein structure prediction and model quality assessment. Drug Discov Today 2009;14(7-8):386–93. https://doi.org/10.1016/j.drudis.2008.11.010.

[3] Kryshtafovych A, Barbato A, Fidelis K, Monastyrskyy B, Schwede T, et al. Assessment of the assessment: evaluation of the model quality estimates in CASP10. Proteins 2014;82(Suppl 2):112–26. https://doi.org/10.1002/prot.24347.

[4] Kryshtafovych A, Barbato A, Monastyrskyy B, Fidelis K, Schwede T, et al. Methods of model accuracy estimation can help selecting the best models from decoy sets: assessment of model accuracy estimations in CASP11. Proteins 2016;84(Suppl 1):349–69. https://doi.org/10.1002/prot.24919.

[5] Jauch R, Yeo HC, Kolatkar PR, Clarke ND. Assessment of CASP7 structure predictions for template free targets. Proteins 2007;69(Suppl 8):57–67. https://doi.org/10.1002/prot.21771.

[6] Haas J, Barbato A, Behringer D, Studer G, Roth S, et al. Continuous automated model evaluation (CAMEO) complementing the critical assessment of structure prediction in CASP12. Proteins 2018;86(Suppl 1):387–98. https://doi.org/10.1002/prot.25431.

[7] Studer G, Tauriello G, Schwede T. Assessment of the assessment-all about complexes. Proteins 2023;91(12):1850–60. https://doi.org/10.1002/prot.26612.

[8] Wang Z, Eickholt J, Cheng J. MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. Bioinformatics 2010; 26(7):882–8. https://doi.org/10.1093/bioinformatics/btq058.

[9] Cheng J, Wang Z, Tegge AN, Eickholt J. Prediction of global and local quality of CASP8 models by MULTICOM series. Proteins 2009;77(Suppl 9):181–4. https://doi.org/10.1002/prot.22487.

[10] Wang W, Li Z, Wang J, Xu D, Shang Y. PSICA: a fast and accurate web service for protein model quality analysis. Nucleic Acids Res 2019;47(W1):W443–50. https://doi.org/10.1093/nar/gkz402.

[11] Wang W, Wang J, Li Z, Xu D, Shang Y. MUfoldQA_G: High-accuracy protein model QA via retraining and transformation. Comput Struct Biotechnol J 2021; 19:6282–90. https://doi.org/10.1016/j.csbj.2021.11.021.

[12] McGuffin LJ. Benchmarking consensus model quality assessment for protein fold recognition. BMC Bioinforma 2007;8:345. https://doi.org/10.1186/1471-2105-8-345.

[13] McGuffin LJ. Prediction of global and local model quality in CASP8 using the ModFOLD server. Proteins 2009;77(Suppl 9):185–90. https://doi.org/10.1002/prot.22491.

[14] Alapati R., Bhattacharya D. (2018, August) clustQ: Efficient protein decoy clustering using superposition-free weighted internal distance comparisons. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics pp. 307–314. https://doi.org/10.1145/3233547.3233570.

[15] Lundström J, Rychlewski L, Bujnicki J, Elofsson A. Pcons: a neural-network-based consensus predictor that improves fold recognition. Protein Sci: a Publ Protein Soc 2001;10(11):2354–62. https://doi.org/10.1110/ps.08501.

[16] Wang Z, Eickholt J, Cheng J. APOLLO: a quality assessment service for single and multiple protein models. Bioinformatics 2011;27(12):1715–6. https://doi.org/10.1093/bioinformatics/btr268.

[17] McGuffin LJ, Aldowsari FMF, Alharbi SMA, Adiyaman R. ModFOLD8: accurate global and local quality estimates for 3D protein models. Nucleic Acids Res 2021; 49(W1):W425–30. https://doi.org/10.1093/nar/gkab321.

[18] McGuffin LJ. The ModFOLD server for the quality assessment of protein structural models. Bioinforma (Oxf, Engl) 2008;24(4):586–7. https://doi.org/10.1093/bioinformatics/btn014.

[19] Maghrabi AHA, McGuffin LJ. Estimating the quality of 3D protein models using the ModFOLD7 server. Methods Mol Biol 2020;2165:69–81. https://doi.org/10.1007/978-1-0716-0708-4_4.

[20] Studer G, Rempfer C, Waterhouse AM, Gumienny R, Haas J, et al. QMEANDisCo-distance constraints applied on model quality estimation. Bioinformatics 2020;36 (6):1765–71. https://doi.org/10.1093/bioinformatics/btz828.

[21] Olechnovič K, Venclovas Č. VoroMQA web server for assessing three-dimensional structures of proteins and protein complexes. Nucleic Acids Res 2019;47(W1): W437–42. https://doi.org/10.1093/nar/gkz367.

[22] Igashov I, Olechnovič K, Kadukova M, Venclovas Č, Grudinin S. VoroCNN: deep convolutional neural network built on 3D Voronoi tessellation of protein structures. Bioinformatics 2021;37(16):2332–9. https://doi.org/10.1093/bioinformatics/btab118.

[23] Olechnovič K, Venclovas Č. VoroMQA: assessment of protein structure quality using interatomic contact areas. Proteins 2017;85(6):1131–45. https://doi.org/10.1002/prot.25278.

[24] Guo SS, Liu J, Zhou XG, Zhang GJ. 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.

[25] Liu J, Zhao K, Zhang G. Improved model quality assessment using sequence and structural information by enhanced deep neural networks. Brief Bioinforma 2023; 24(1):bbac507. https://doi.org/10.1093/bib/bbac507.

[26] Baldassarre F, Menéndez Hurtado D, Elofsson A, Azizpour H. GraphQA: protein model quality assessment using graph convolutional networks. Bioinformatics 2021;37(3):360–6. https://doi.org/10.1093/bioinformatics/btaa714.

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

[28] Hiranuma N, Park H, Baek M, Anishchenko I, Dauparas J, et al. Improved protein structure refinement guided by deep learning based accuracy estimation. Nat Commun 2021;12(1):1340. https://doi.org/10.1038/s41467-021-21511-x.

[29] Ye L, Wu P, Peng Z, Gao J, Liu J, et al. Improved estimation of model quality using predicted inter-residue distance. Bioinformatics 2021;37(21):3752–9. https://doi.org/10.1093/bioinformatics/btab632.

[30] 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(Suppl 1):i285–91. https://doi.org/10.1093/bioinformatics/btaa455.

[31] Pagès G, Charmettant B, Grudinin S. Protein model quality assessment using 3D oriented convolutional neural networks. Bioinformatics 2019;35(18):3313–9. https://doi.org/10.1093/bioinformatics/btz122.

[32] Conover M, Staples M, Si D, Sun M, Cao R. AngularQA: protein model quality assessment with LSTM networks. Comput Math Biophys 2019;7(1):1–9. https://doi.org/10.1515/cmb-2019-0001.

[33] Uziela K, Menéndez Hurtado D, Shu N, Wallner B, Elofsson A. ProQ3D: improved model quality assessments using deep learning. Bioinformatics 2017;33(10): 1578–80. https://doi.org/10.1093/bioinformatics/btw819.

[34] Uziela K, Wallner B. ProQ2: estimation of model accuracy implemented in Rosetta. Bioinformatics 2016;32(9):1411–3. https://doi.org/10.1093/bioinformatics/btv767.

[35] Hurtado DM, Uziela K, Elofsson A. Deep transfer learning in the assessment of the quality of protein models. arXiv preprint 2018. https://doi.org/10.48550/arXiv.1804.06281.

[36] Ray A, Lindahl E, Wallner B. Improved model quality assessment using ProQ2. BMC Bioinforma 2012;13:224. https://doi.org/10.1186/1471-2105-13-224.

[37] Basu S, Wallner B. Finding correct protein-protein docking models using ProQDock. Bioinformatics 2016;32(12):i262–70. https://doi.org/10.1093/bioinformatics/btw257.

[38] Milner JL, Wood JM. Insertion proQ220::Tn5 alters regulation of proline porter II, a transporter of proline and glycine betaine in Escherichia coli. J Bacteriol 1989; 171(2):947–51. https://doi.org/10.1128/jb.171.2.947-951.1989.

[39] Wallner B, Elofsson A. Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Sci: a Publ Protein Soc 2006;15(4):900–13. https://doi.org/10.1110/ps.051799606.

[40] 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.

[41] Derevyanko G, Grudinin S, Bengio Y, Lamoureux G. Deep convolutional networks for quality assessment of protein folds. Bioinformatics 2018;34(23):4046–53. https://doi.org/10.1093/bioinformatics/bty494.

[42] Cao R, Adhikari B, Bhattacharya D, Sun M, Hou J, et al. QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. Bioinformatics 2017;33(4):586–8. https://doi.org/10.1093/bioinformatics/btw694.

[43] Manavalan B, Lee J. SVMQA: support-vector-machine-based protein single-model quality assessment. Bioinformatics 2017;33(16):2496–503. https://doi.org/10.1093/bioinformatics/btx222.

[44] Cao R, Bhattacharya D, Hou J, Cheng J. DeepQA: improving the estimation of single protein model quality with deep belief networks. BMC Bioinforma 2016;17 (1):495. https://doi.org/10.1186/s12859-016-1405-y.

[45] Benkert P, Tosatto SC, Schomburg D. QMEAN: a comprehensive scoring function for model quality assessment. Proteins 2008;71(1):261–77. https://doi.org/10.1002/prot.21715.

[46] Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins 1993;17(4):355–62. https://doi.org/10.1002/prot.340170404.

[47] Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 2007;35 (Web Server issue):W407–10. https://doi.org/10.1093/nar/gkm290.

[48] Cheng J, Choe MH, Elofsson A, Han KS, Hou J, et al. Estimation of model accuracy in CASP13. Proteins 2019;87(12):1361–77. https://doi.org/10.1002/prot.25767.

[49] Elofsson A, Joo K, Keasar C, Lee J, Maghrabi AHA, et al. Methods for estimation of model accuracy in CASP12. Proteins 2018;86(Suppl 1):361–73. https://doi.org/10.1002/prot.25395.

[50] Liu D, Zhang B, Liu J, Li H, Song L, et al. Assessing protein model quality based on deep graph coupled networks using protein language model. Brief Bioinforma 2023;25(1):bbad420. https://doi.org/10.1093/bib/bbad420.

[51] Fowler NJ, Williamson MP. The accuracy of protein structures in solution determined by AlphaFold and NMR. Structures 2022;30(7):925–33. https://doi.org/10.1016/j.str.2022.04.005.

[52] Cramer P. AlphaFold2 and the future of structural biology. Nat Struct Mol Biol 2021;28(9):704–5. https://doi.org/10.1038/s41594-021-00650-1.

[53] Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, et al. CAPRI: a critical assessment of PRedicted interactions. Proteins 2003;52(1):2–9. https://doi.org/10.1002/prot.10381.

[54] Lensink MF, Brysbaert G, Raouraoua N, Bates PA, Giulini M, et al. Impact of AlphaFold on structure prediction of protein complexes: the CASP15-CAPRI experiment. Proteins 2023;91(12):1658–83. https://doi.org/10.1002/prot.26609.

[55] Lensink MF, Velankar S, Kryshtafovych A, Huang SY, Schneidman-Duhovny D, et al. Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: a CASP-CAPRI experiment. Proteins 2016;84 (Suppl 1):323–48. https://doi.org/10.1002/prot.25007.

[56] Lensink MF, Velankar S, Baek M, Heo L, Seok C, et al. The challenge of modeling protein assemblies: the CASP12-CAPRI experiment. Proteins 2018;86(Suppl 1): 257–73. https://doi.org/10.1002/prot.25419.

[57] Lensink MF, Brysbaert G, Mauri T, Nadzirin N, Velankar S, et al. Prediction of protein assemblies, the next frontier: the CASP14-CAPRI experiment. Proteins 2021;89(12):1800–23. https://doi.org/10.1002/prot.26222.

[58] Lensink MF, Brysbaert G, Nadzirin N, Velankar S, Chaleil RAG, et al. Blind prediction of homo- and hetero-protein complexes: the CASP13-CAPRI experiment. Proteins 2019;87(12):1200–21. https://doi.org/10.1002/prot.25838.

[59] Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31(13):3370–4. https://doi.org/10.1093/nar/gkg571.

[60] Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57(4):702–10. https://doi.org/10.1002/prot.20264.

[61] 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(12):1889–902. https://doi.org/10.1002/prot.26542.

[62] 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.

[63] Bertoni M, Kiefer F, Biasini M, Bordoli L, Schwede T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by

homology. Sci Rep 2017;7(1):10480. https://doi.org/10.1038/s41598-017-09654-8.

[64] Edmunds NS, Alharbi SMA, Genc AG, Adiyaman R, McGuffin LJ. Estimation of model accuracy in CASP15 using the ModFOLDdock server. Proteins 2023;91 (12):1871–8. https://doi.org/10.1002/prot.26532.

[65] 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(21):2722–8. https://doi.org/10.1093/bioinformatics/btt473.

[66] Olechnovič K, Kulberkytė E, Venclovas C. CAD-score: a new contact area difference-based function for evaluation of protein structural models. Proteins 2013;81(1):149–62. https://doi.org/10.1002/prot.24172.

[67] Liu J, Liu D, Zhang GJ. DeepUMQA3: a web server for accurate assessment of interface residue accuracy in protein complexes. Bioinformatics 2023;39(10): btad591. https://doi.org/10.1093/bioinformatics/btad591.

[68] 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(3):273–90. https://doi.org/10.1037/met0000079.

[69] Muschelli J. ROC and AUC with a binary predictor: a potentially misleading metric. J Classif 2020;37(3):696–708. https://doi.org/10.1007/s00357-019-09345-1.

[70] Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999;Suppl 3:22–9. https://doi.org/10.1002/(sici)1097-0134(1999)37:3+<22::aid-prot5>3.3.co;2-n.

[71] Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? Bioinforma (Oxf, Engl) 2010;26(7):889–95. https://doi.org/10.1093/bioinformatics/btq066.

[72] Shor B, Schneidman-Duhovny D. CombFold: predicting structures of large protein assemblies using a combinatorial assembly algorithm and AlphaFold2. Nat Methods 2024;21(3):477–87. https://doi.org/10.1038/s41592-024-02174-0.

[73] Won J, Baek M, Monastyrskyy B, Kryshtafovych A, Seok C. Assessment of protein model structure accuracy estimation in CASP13: challenges in the era of deep learning. Proteins 2019;87(12):1351–60. https://doi.org/10.1002/prot.25804.

[74] Olechnovič K, Valančauskas L, Dapkūnas J, Venclovas Č. Prediction of protein assemblies by structure sampling followed by interface-focused scoring. Proteins 2023;91(12):1724–33. https://doi.org/10.1002/prot.26569.

[75] Liu J, Guo Z, Wu T, Roy RS, Quadir F, et al. Enhancing alphafold-multimer-based protein complex structure prediction with MULTICOM in CASP15. Commun Biol 2023;6(1):1140. https://doi.org/10.1038/s42003-023-05525-3.

[76] Olechnovič K, Venclovas Č. VoroIF-GNN: voronoi tessellation-derived protein-protein interface assessment using a graph neural network. Proteins 2023;91(12): 1879–88. https://doi.org/10.1002/prot.26554.

[77] He G, Liu J, Liu D, Zhang G. GraphGPSM: a global scoring model for protein structure using graph neural networks. Brief Bioinforma 2023;24(4):bbad219. https://doi.org/10.1093/bib/bbad219.

[78] 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(12): 1861–70. https://doi.org/10.1002/prot.26564.

[79] Ngounou Wetie AG, Sokolowska I, Woods AG, Roy U, Deinhardt K, et al. Protein-protein interactions: switch from classical methods to proteomics and bioinformatics-based approaches. Cell Mol life Sci 2014;71(2):205–28. https://doi.org/10.1007/s00018-013-1333-1.

[80] Lensink MF, Wodak SJ. Docking, scoring, and affinity prediction in CAPRI. Proteins 2013;81(12):2082–95. https://doi.org/10.1002/prot.24428.

[81] Lensink MF, Méndez R, Wodak SJ. Docking and scoring protein complexes: CAPRI 3rd Edition. Proteins 2007;69(4):704–18. https://doi.org/10.1002/prot.21804.

[82] 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.

[83] Zhu W, Shenoy A, Kundrotas P, Elofsson A. Evaluation of alphafold-multimer prediction on multi-chain protein complexes. Bioinformatics 2023;39(7): btad424. https://doi.org/10.1093/bioinformatics/btad424.

[84] 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.

[85] 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.

[86] Mohseni Behbahani Y, Crouzet S, Laine E, Carbone A. Deep Local Analysis evaluates protein docking conformations with locally oriented cubes. Bioinformatics 2022;38(19):4505–12. https://doi.org/10.1093/bioinformatics/btac551.

[87] McGuffin LJ, Edmunds NS, Genc AG, Alharbi SMA, Salehe BR, et al. Prediction of protein structures, functions and interactions using the IntFOLD7, MultiFOLD and ModFOLDdock servers. Nucleic Acids Res 2023;51(W1):W274–80. https://doi.org/10.1093/nar/gkad297.

[88] Notredame C. Recent evolutions of multiple sequence alignment algorithms. PLOS Comput Biol 2007;3(8):e123. https://doi.org/10.1371/journal.pcbi.0030123.

[89] Kryshtafovych A, Antczak M, Szachniuk M, Zok T, Kretsch RC, et al. New prediction categories in CASP15. Proteins 2023;91(12):1550–7. https://doi.org/10.1002/prot.26515.

[90] 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(6637):1123–30. https://doi.org/10.1126/science.ade2574.

[91] 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.

[92] Pereira J, Simpkin AJ, Hartmann MD, Rigden DJ, Keegan RM, et al. High-accuracy protein structure prediction in CASP14. Proteins 2021;89(12):1687–99. https://doi.org/10.1002/prot.26171.

[93] Maheshwari S, Brylinski M. Template-based identification of protein-protein interfaces using eFindSitePPI. Methods 2016;93:64–71. https://doi.org/10.1016/j.ymeth.2015.07.017.

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

[95] Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 2017;35(11):1026–8. https://doi.org/10.1038/nbt.3988.

[96] Zhang J, Wang Q, Barz B, He Z, Kosztin I, et al. MUFOLD: a new solution for protein 3D structure prediction. Proteins 2010;78(5):1137–52. https://doi.org/10.1002/prot.22634.

[97] Dapkūnas J, Olechnovič K, Venclovas Č. Modeling of protein complexes in CASP14 with emphasis on the interaction interface prediction. Proteins 2021;89 (12):1834–43. https://doi.org/10.1002/prot.26167.

[98] Ozden B, Kryshtafovych A, Karaca E. The impact of AI-based modeling on the accuracy of protein assembly prediction: insights from CASP15. Proteins 2023;91 (12):1636–57. https://doi.org/10.1002/prot.26598.

[99] Zhao K, Xia Y, Zhang F, Zhou X, Li SZ, et al. Protein structure and folding pathway prediction based on remote homologs recognition using PAthreader. Commun Biol 2023;6(1):243. https://doi.org/10.1038/s42003-023-04605-8.

[100] Simpkin AJ, Mesdaghi S, Sánchez Rodríguez F, Elliott L, Murphy DL, et al. Tertiary structure assessment at CASP15. Proteins 2023;91(12):1616–35. https://doi.org/10.1002/prot.26593.

[101] Ruffolo JA, Chu LS, Mahajan SP, Gray JJ. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. Nat Commun 2023;14(1):2389. https://doi.org/10.1038/s41467-023-38063-x.

[102] Mitchell LS, Colwell LJ. Comparative analysis of nanobody sequence and structure data. Proteins 2018;86(7):697–706. https://doi.org/10.1002/prot.25497.

[103] Leem J, Mitchell LS, Farmery JHR, Barton J, Galson JD. Deciphering the language of antibodies using self-supervised learning. Patterns 2022;3(7):100513. https://doi.org/10.1016/j.patter.2022.100513.

[104] Wang D., Ye F., Zhou H.2023 On pre-trained language models for antibody. arXiv preprint arXiv:230112112. https://doi.org/10.48550/arXiv.2301.12112.

[105] Janson G, Valdes-Garcia G, Heo L, Feig M. Direct generation of protein conformational ensembles via machine learning. Nat Commun 2023;14(1):774. https://doi.org/10.1038/s41467-023-36443-x.

[106] Deep A, Gu Y, Gao YQ, Ego KM, Herzik Jr MA, et al. The SMC-family Wadjet complex protects bacteria from plasmid transformation by recognition and cleavage of closed-circular DNA. Mol Cell 2022;82(21):4145–59. https://doi.org/10.1016/j.molcel.2022.09.008.