Assessing scoring metrics for AlphaFold2 and AlphaFold3 protein complex predictions

Luca R. Genz[1,2,3], Sanjana Nair[1,3] and Maya Topf*[1,3,4]

[1]Leibniz-Institut für Virologie (LIV), [2]Universität Hamburg, [3]Centre for Structural Systems Biology (CSSB) and [4]Universitätsklinikum Hamburg Eppendorf (UKE), Hamburg, Germany

*To whom correspondence should be addressed.

Correspondence: Maya Topf; maya.topf@cssb-hamburg.de; +49 040 8998 87660

1. Abstract

Abstract: The recent breakthroughs in AI-driven protein structure prediction have revolutionized structural biology, unlocking new possibilities to model complex biomolecular interactions. We evaluated widely-used scoring metrics for assessing such models predicted by ColabFold with templates, ColabFold without templates, and AlphaFold3. We benchmarked the optimal cutoffs for these assessment scores using a set of 325 heterodimeric high-resolution protein structures and their predictions. Our results show that ColabFold with templates and AlphaFold3 perform similarly and both outperform ColabFold without templates. Furthermore, interface-specific scores are found to be more reliable for evaluating protein complex predictions compared to the corresponding global scores. Notably, ipTM and *model confidence* achieve the best performance in distinguishing correct from incorrect predictions. Based on our results, we developed a weighted combined score, C2Qscore, to improve model quality assessment, which was used to analyze dimers from large assemblies solved by cryoEM. This revealed potential limitations of the

assessment scores when multiple configurations of heterodimers are possible. C2Qscore has been integrated as a tool into our ChimeraX plug-in PICKLUSTER v2.0, facilitating interactive access to the scoring metrics. The plug-in is freely available for download from the ChimeraX Toolshed and https://gitlab.com/topf-lab/pickluster-v2.0.git. This study provides insights into the strengths and weaknesses of current scores and offers guidance for improving protein complex model assessment.

Keywords: Protein complex modeling; Interface scoring; Scoring metrics; AlphaFold3; ColabFold, protein structure prediction; ChimeraX plug-in

Impact of this work: This study evaluates the effectiveness of current scoring metrics used to assess computationally generated structural models of protein complexes, which are crucial for understanding cellular functions. By analyzing the performance of these metrics and comparing different prediction methods, we aim to enhance the reliability and accuracy of such models. Our findings provide valuable insights for researchers utilizing computational approaches to study biological processes and have potential applications in biomedical research.

2. Introduction

Protein complexes are fundamental to various cellular processes, performing diverse functions in vivo (Schulz and Schirmer 2013). To understand the mechanisms of protein complex formation and macromolecular recognition, the precise identification and characterization of protein-protein interfaces is a crucial step (Xue et al. 2015). High-resolution structural data obtained from experimental methods such as X-ray crystallography and cryogenic-electron microscopy (cryo-EM) offer insights into these interfaces (Smyth and Martin 2000; Namba and Makino 2022). However, these are resource-intensive techniques and present challenges for large-scale structure determination (Xue et al. 2015). Recent advancements in AI-based protein structure prediction, particularly AlphaFold2, have mitigated these challenges by enabling large-scale structure predictions sometimes even with accuracies comparable to those of experimental methods (Jumper et al. 2021; Evans et al. 2022). This progress was recognized with the 2024 Nobel Prize in Chemistry. AlphaFold2's successor, AlphaFold3, further advanced the field by allowing the prediction of protein complexes with lipids, nucleic acids and other ligands (Abramson et al. 2024).

Beyond predicting the structures of protein complexes, it is equally important to assess the accuracy of these predictions to ensure confidence in the models which is crucial for downstream analyses, such as functional studies, protein engineering, and drug design (Baker and Sali 2001). However, evaluating the quality of predictions without a reference structure is challenging. The progress in the protein structure prediction field has led to the development of various methods for accessing model accuracy, with increasing emphasis on evaluating protein complexes (Studer et al. 2023). CASP (Critical Assessment of Structure Prediction), a community-wide experiment designed to track advancements in the structure prediction field, introduced the Estimation of

Model Accuracy (EMA) category already in CASP7 to address this need (Cozzetto et al. 2007).

To date, numerous prediction-based evaluation metrics have been proposed, most of which are machine learning-based (Studer et al. 2023 Oct 19). The graph neural network (GNN)-based scoring method, VoroIF-GNN (VoroIF) (Olechnovič and Venclovas 2023), emerged as one of the top performing scores in assessing interface quality in the CASP15 EMA iteration, ranking alongside ModFOLDdockR and ModFOLDdock (Edmunds et al. 2023) and VoroMQA (Olechnovič and Venclovas 2019). VoroIF uses Voronoi tessellation to derive interface graphs, providing a detailed, contact-based accuracy estimate for the entire interface. The VoroIF interface-specific score is calculated as the average score of interface residues (Average-VoroIF-GNN-residue-pCAD score), which is then adjusted for structural completeness based on the ratio of atoms in the model to atoms in the target.

Another commonly used metric for evaluating the quality of protein structure predictions is the predicted DockQ score (pDockQ) (Bryant et al. 2022). This score is derived by calculating the number of interfacial contacts and the average quality of interacting residues, which are then fitted to a sigmoid function of the DockQ score (Basu and Wallner 2016). Its more recent iteration, pDockQ2 (Zhu et al. 2023), was specifically developed for the assessment of multimeric protein complexes.

To our knowledge, to date only Yin et al. (Yin et al. 2022) have conducted a systematic assessment of scores that evaluate AlphaFold2 multimeric predictions, using a relatively small benchmark set of 152 heterodimeric protein complexes. In their analysis, they compared the performance of AlphaFold2 to the protein-docking algorithm ZDOCK (Pierce et al. 2011). Evaluated metrics included the predicted Local Distance Difference Test (pLDDT) (Jumper et al. 2021), an interface-specific version

100 of pLDDT, predicted Template Modeling score (pTM) (Jumper et al. 2021), an
101 interface-specific Pairwise Alignment Error (PAE) (Evans et al. 2022), among other
102 metrics, and these were compared against the interface Root Mean Square Deviation
103 (I-RMSD) (Lensink and Wodak 2013; Basu and Wallner 2016). Furthermore, most of
104 the current evaluation metrics, including pDockQ, VoroIF, and others, have been
105 applied in isolation and were evaluated prior to the development of AlphaFold3.
106 In this study, we focused on the quality of the interface, by providing a comprehensive
107 comparison and evaluation of commonly used scores for assessing predictions of
108 heterodimeric protein complexes. We evaluated predictions from both
109 AlphaFold2/ColabFold (Mirdita et al. 2022) (with and without templates) and
110 AlphaFold3. We compared different scores, with DockQ as ground truth, including:
111 interface predicted Local Distance Difference Test (ipLDDT) (Genz et al. 2023), pTM,
112 interface pTM (ipTM) (Evans et al. 2022), *model confidence* (Evans et al. 2022),
113 interface PAE (iPAE) (Evans et al. 2022; Genz et al. 2023), pDockQ2 (Zhu et al. 2023),
114 and VoroIF. The individual scores were merged into one score, trained on predictions
115 for 325 heterodimeric high-resolution structures from the Protein Data Bank (PDB) and
116 tested on an independent dataset of predictions for X-ray crystallographic structures
117 and dimers from larger assemblies derived from cryoEM. The score has been
118 integrated as a tool into our ChimeraX plug-in PICKLUSTER v2.0 (Genz et al. 2023),
119 providing users with interactive access to scoring metrics as a valuable analytical tool.

120 ## 3. Results

121 ### 3.1. Benchmarking the prediction accuracy of ColabFold with templates,
122 ### ColabFold template free and AlphaFold3

123 To evaluate the capability of ColabFold and AlphaFold3 in generating high-quality
124 predictions of structural models, a comparative analysis was performed using a

benchmark set. We focused exclusively on heterodimeric complexes rather than homodimeric ones, as AlphaFold2 generally performs better on homomeric interfaces (Evans et al. 2022; Zhu et al. 2023). By selecting heterodimers, we aimed to introduce greater diversity and a more challenging evaluation setting. Starting with 1,381 complexes, a filtering process (See Methods, Suppl. Fig. 1) reduced the number of target structures to 397. Among these, 72 cases were identified in which the biological assembly (BA) assigned in the PDB is different from the asymmetric unit (AU), resulting in predictions resembling only one of the two. This caused issues during evaluation with DockQ, as DockQ was used to align the AU with the predicted model, leading to artificially low scores that do not reflect the prediction quality in these specific cases. To address this, only templates with the assigned dimeric BA identical to the AU were included in the benchmark set, resulting in a final set of 325 target structures. Predictions were generated using ColabFold with templates (CF-T) and with the template-free (CF-F) option, and the AlphaFold3 web server (AF3). All predictions were performed with three recycles followed by relaxation (Hornak et al. 2006), producing five predictions per target structure. DockQ and individual prediction-based scores were calculated for a total of 1,612 models in the CF-T and 1,614 models in the CF-F set. In the AF3 set, scores were obtained for 1,595 models.

Using CAPRI (Critical Assessment of PRediction of Interactions) criteria for DockQ ("DockQ classification", Methods) revealed that CF-T (52.5%) and AF3 (50.3%) had the highest proportion of 'high' quality models (DockQ > 0.8), making it the largest fraction for both prediction methods (Fig. 1A).

Of the three prediction methods, the CF-F set contained the lowest proportion of 'high' quality models at 34.7% with the largest fraction consisting of models of 'medium' quality.

Analyzing the models classified as ‘incorrect’ using DockQ (DockQ ≤ 0.23) showed that AF3 had the lowest percentage of incorrect models at 13.5%. In contrast, the CF-T and CF-F sets exhibited significantly higher proportions of incorrect models, at 17.5% and 22.2%, respectively.

Further investigation of the composition of the ‘incorrect’ fraction revealed that AF3 showed the highest percentage of cases where all five models of the prediction were classified as incorrect (83.7%) (Fig. 1B) CF-F revealed a significantly smaller proportion of cases where all five models of the prediction were classified as incorrect (75.4%) and the CF-T set had the lowest percentage of those (69.5%).

Considering only the ‘high’ quality fraction, CF-T predictions appear to be the most accurate. However, if models classified as ‘acceptable’, ‘medium’ and ‘high’ quality according to the CAPRI criteria are all considered correct (as opposed to the ‘incorrect’ fraction), then AF3 predictions are superior, as they exhibit the lowest number of incorrect models. This detailed analysis of the quality of the predicted models suggests AlphaFold3 and ColabFold with templates as the most reliable predictors. Additionally, it demonstrates that compared to ColabFold, AlphaFold3 predictions are more consistent in regards to whether a model is correct or incorrect. This is indicated by the significantly higher proportion of cases where all five models in a prediction were classified as incorrect and the lowest proportion where only one model is incorrect.

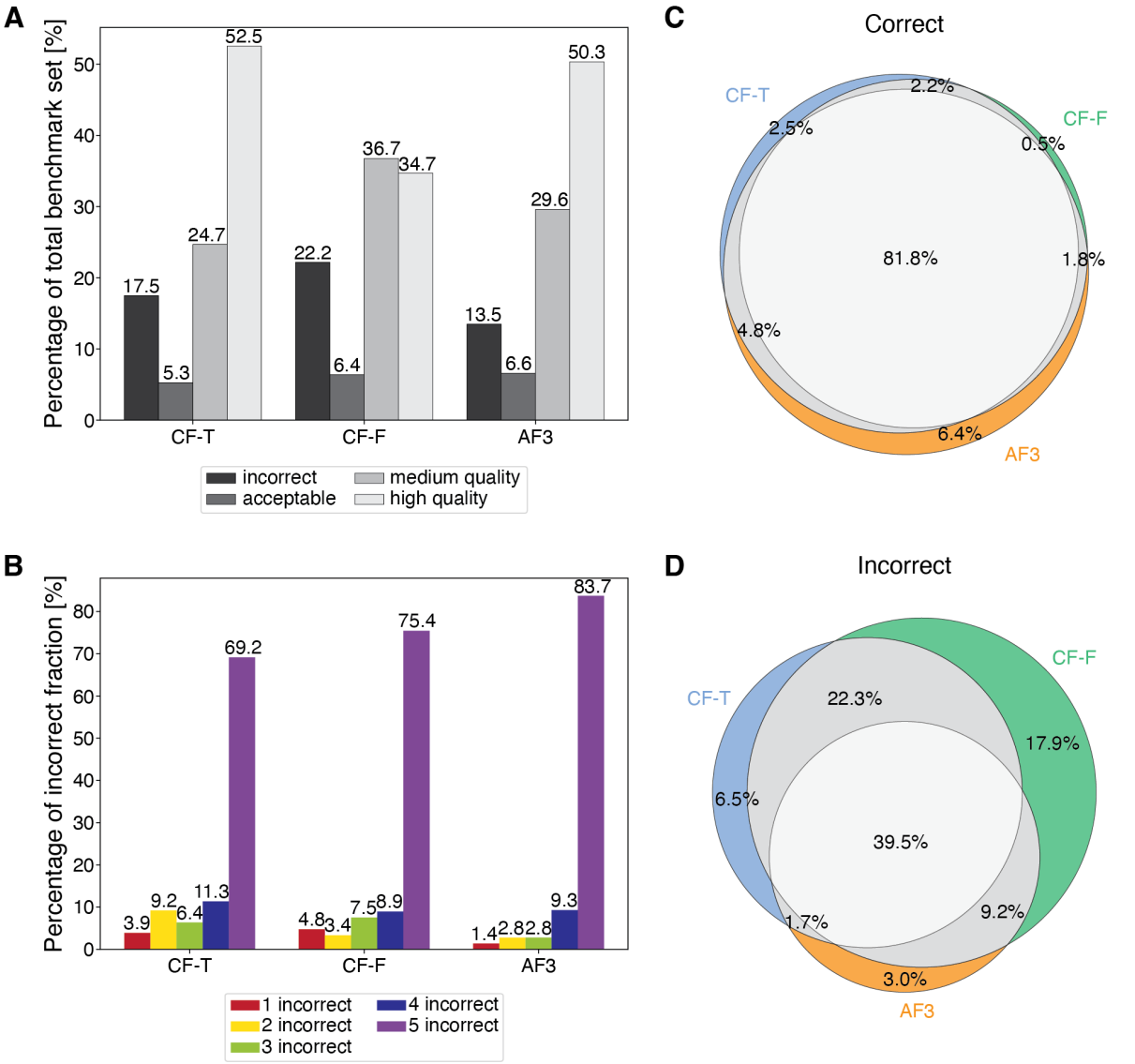


Fig. 1. Quality distributions and overlap of DockQ classifications (5 levels) across models predicted by CF-T, CF-F, and AF3. **A:** Quality assessment of the three sets based on DockQ classification criteria. **B:** Breakdown of the incorrect fraction from A, categorizing cases where all five models of a template are classified as incorrect, followed by cases with 4, 3, 2, and those with only one model classified as incorrect. **C:** Venn diagram of the 'correct' fraction ('acceptable', 'medium', 'high' are combined). The fraction of only CF-T predicting models of a specific quality are shown in blue, CF-F in green, and AF3 in orange. When two prediction methods agree, the fraction is represented in gray, and when all three prediction methods agree, it is shown in light gray. **D:** Venn diagram of the 'incorrect' fraction. Colors are as in (C).

180 Moreover, the consensus of prediction quality across the three datasets (CF-T, CF-F
181 and AF3) was analyzed by comparing models of the same rank (Fig. 1C, D, Suppl.
182 Fig. 2). When the 'high', 'medium' and 'acceptable' quality fractions are combined
183 (considered 'correct' based on CAPRI criteria), the models predicted by the three
184 different methods are comparable in terms of quality in 81.8% of cases (Venn diagram
185 in Fig. 1C). The highest proportion (6.4%) of uniquely correct predictions was observed
186 for AF3, suggesting that running AlphaFold3 may enhance the likelihood of generating
187 reliable models. However, Venn diagrams illustrating just high-quality models (based
188 on CAPRI criteria) reveal that only 50.3% of the cases were consistent across all three
189 datasets (Suppl. Fig. 2). Notably, for high-quality models only, the highest proportion
190 (14.9%) of predictions when only one method was able to achieve such results was
191 observed for CF-T. This suggests that using ColabFold with templates may increase
192 the likelihood of generating high-quality models. The Venn diagram shows that CF-F
193 had the highest proportion of incorrect predictions (17.9%) based on CAPRI criteria
194 (Fig. 1D), consistent with its highest proportion of incorrect models in the benchmark
195 set (Fig. 1A). The acceptable fraction showed the lowest agreement (17.6%) between
196 the three prediction methods (Suppl. Fig. 2), suggesting that in cases where
197 predictions are near the threshold for incorrect predictions (DockQ ≤ 0.23), running all
198 three methods and comparing results may yield the most optimal outcome.

199 **3.2. Distributions and correlations of individual prediction-based scores**

200 In order to evaluate the effectiveness of the raw assessment scores (ipLDDT, iPAE,
201 pTM, ipTM, *model confidence*, pDockQ2 and VoroIF, Suppl. Tab. 1) in predicting the
202 quality of the models, these scores were calculated for each model in the three distinct
203 datasets: CF-T, CF-F and AF3. Given the variability in defining the interface based on
204 different distance cutoffs and different types of atoms and positions, the impact of the

distance parameter on ipLDDT and iPAE was investigated (see Methods). The results indicated no substantial difference in interface detection or ipLDDT and iPAE calculations across the various distance parameters (Suppl. Fig. 3). This suggests that the distance parameter has a marginal influence on these scores.

Consequently, a heavy atom distance cutoff of 5 Å was adopted for the definition of the interface and the calculation of the ipLDDT and the iPAE in agreement with the interface threshold used by AlphaFold to define the interface (Evans et al. 2022).

Subsequently, the empirical distributions of the scores and the correlation between the scores were analyzed (Fig. 2, Supp Fig. 4,5). A common trend across these distributions was the predominance of high scores (low scores for iPAE and PAE), suggesting a generally high quality of the prediction sets, consistent with the findings discussed in 3.1.

As expected, *model confidence* exhibited the highest correlation with ipTM across all three datasets, which is the dominant component of the former. Conversely, pDockQ2 demonstrated the lowest correlation with other scores. Overall, the global scores pLDDT and PAE, had a lower correlation with all other scores. In contrast, the interface-specific scores, ipLDDT and iPAE, had a higher correlation with other scores. This could be due to the interface specificity of most of the other scores (except pTM). Overall, the individual prediction-based scores were in good agreement with our observation that these datasets comprised a high proportion of high-quality models. Moreover, the described trends were consistent across all three datasets (CF-T, CF-F and AF3) (Fig. 2, Supp Fig. 4,5).

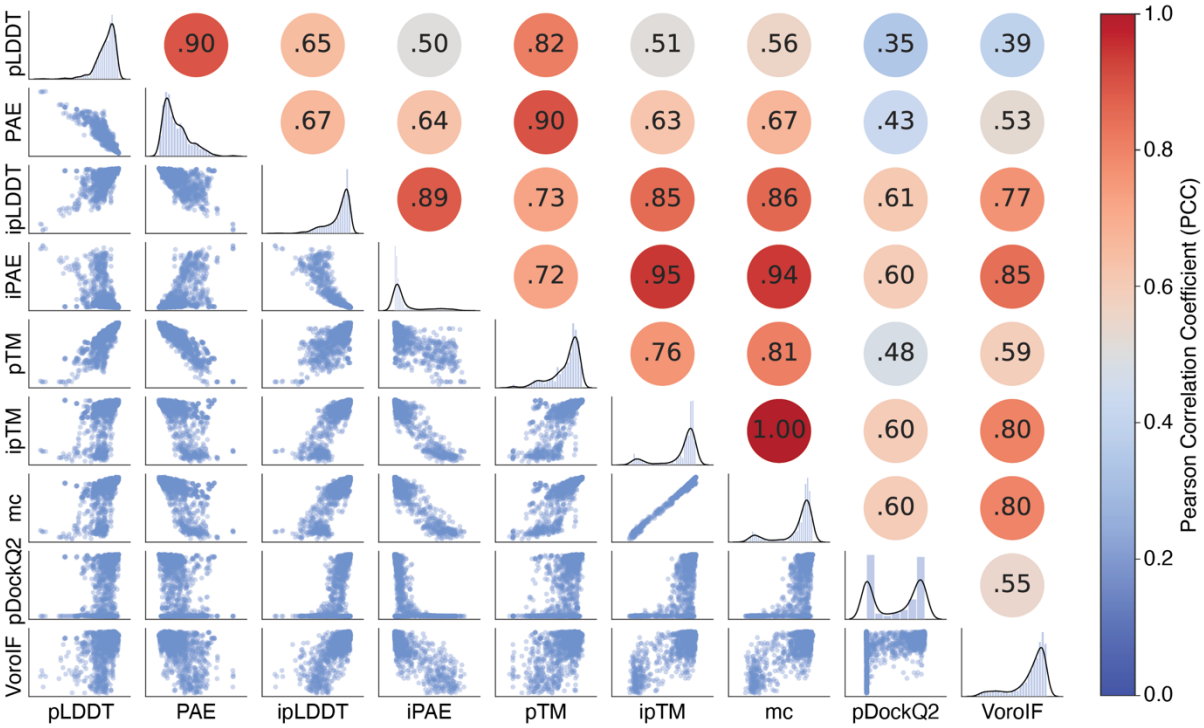


Fig. 2: Distribution and correlation of the individual scores (pLDDT, PAE, ipLDDT, iPAE, pTM, ipTM, *model confidence* (mc), pDockQ2 and VoroIF) for the CF-T dataset. The lower-left section displays the pairwise correlation between the scores. The diagonal shows the distribution of each raw score within the benchmark set. The upper-right section provides the PCC between the scores with red indicating high and blue low PCCs.

3.3. Correlation of individual prediction-based scores with DockQ

To evaluate the effectiveness of individual scores in assessing the quality of the predicted models in the three datasets, PCC between these scores and DockQ were calculated. The results were consistent across CF-T, CF-F, and AF3, with mean PCCs of 0.68, 0.72, and 0.64 (averaged over the individual scores), respectively. However, Wilcoxon signed-rank test showed that only the differences between CF-F and AF3 were statistically significant (p-value of 0.004). These findings suggest that the scores were more effective in evaluating models generated by ColabFold without templates than by AF3. Notably, despite the generally lower quality of models in the CF-F dataset (Fig 1A) compared to CF-T and AF3, the assessment scores performed best for these

243 models (Fig. 3A). ipTM and *model confidence* showed the highest correlations with
244 DockQ across all datasets, recommending these scores as the primary metrics for the
245 protein complex model assessment when only a single score is to be used regardless
246 of the prediction method.
247 Furthermore, the interface-specific scores, ipLDDT and iPAE, outperformed their
248 global versions, pLDDT and PAE, as they exhibited a significantly higher correlation
249 with the DockQ. The third least correlated score was pDockQ2, which remained only
250 mildly correlated with DockQ across all datasets. Fig 3B (i, ii, and iii), provides
251 examples to demonstrate that the ipTM and DockQ indicate high quality models, but
252 pDockQ2 indicates the opposite: for targets PDB: 6UKA (predicted by CF-T) and PDB:
253 5ZZA (predicted by CF-F), although DockQ implies a high-quality model and ipTM is
254 high (0.88, 0.82), pDockQ2 is very low (0.01, 0.01). Finally, a similar situation is
255 observed for a model produced by AF3: for target PDB: 7O4Q, ipTM is high (0.82) and
256 pDockQ2 is significantly lower (0.30).

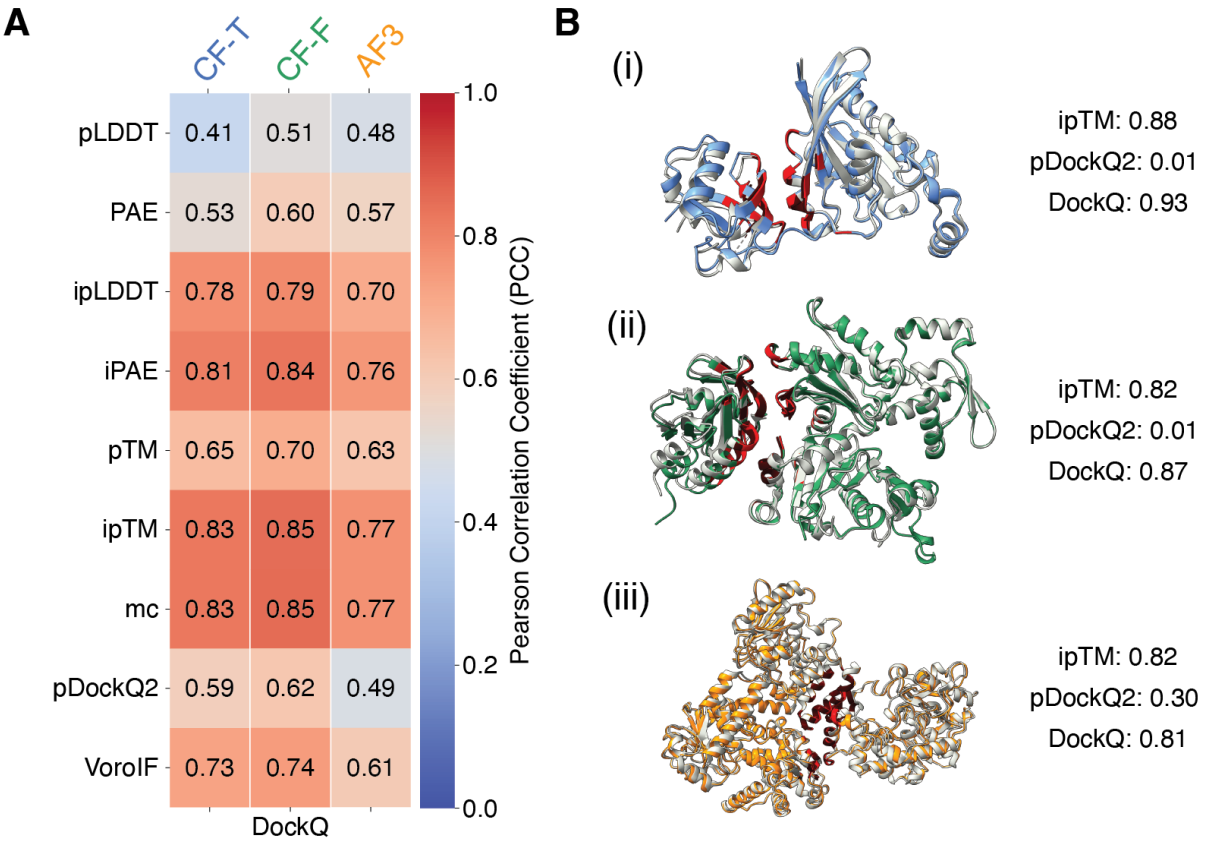


Fig. 3: Quality of the individual scores and representative structural predictions with scores. **A:** PCC of the individual scores (pLDDT, PAE, ipLDDT, iPAE, pTM, ipTM, *model confidence*, pDockQ2 and VoroIF) with DockQ. The color bar is a gradient from blue to red indicating low to high correlation, respectively for CF-T, CF-F and AF3. **B(i):** CF-T prediction (blue, interface: red) and target structure (PDB: 6UKA) (gray, interface: dark red). **B(ii):** CF-F prediction (green, interface: red) and target structure (PDB: 5ZZA) (gray, interface: dark red). **B(iii):** AF3 prediction (orange, interface: red) and target structure (PDB: 7O4Q) (gray, interface: dark red).

3.4. Determinants of high- and low-quality models

The individual scores provide only the absolute values without any established thresholds, thus limiting their application for binary classification of the model as correct or incorrect. To address this limitation, we performed Receiver Operating Characteristic (ROC) curve analyses for each score, allowing the determination of cutoffs based on (Melo and Sali 2007) (see Methods).

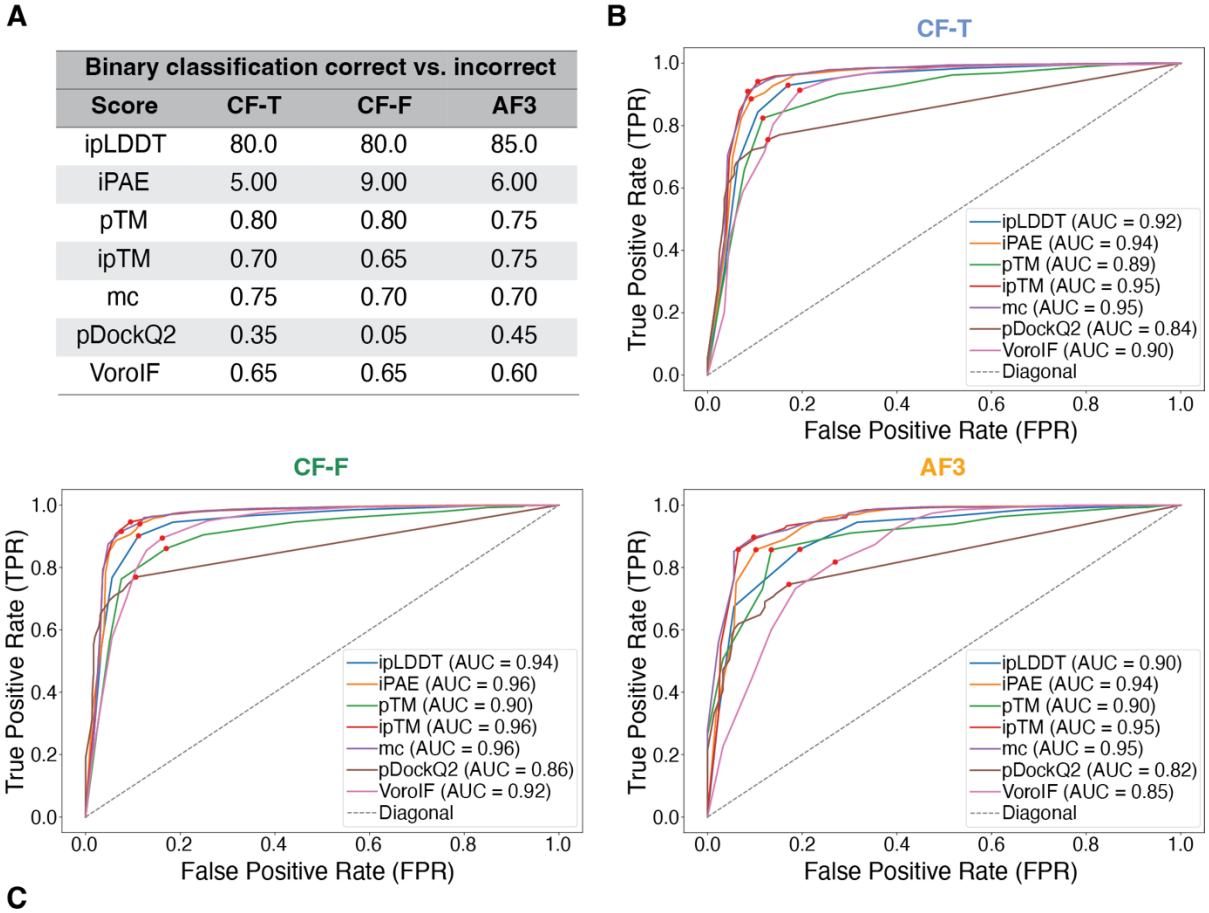
271 For defining the ground truth, models labelled as ‘incorrect’ by DockQ classification
272 were defined as incorrect models, while those falling into the ‘acceptable’, ‘medium’,
273 and ‘high’ DockQ categories were considered as “correct” predictions. Using ROC
274 curves, optimal thresholds were derived for each dataset (Fig. 4A).
275 Additionally, Area Under the Curve (AUC) (Fig. 4B), accuracy and precision (Suppl.
276 Table 2) were calculated to assess the ability of the thresholds to distinguish between
277 models of low and high accuracy as designated by DockQ (see Methods). While the
278 precision values of all three datasets was similarly high, the AUC and accuracy values
279 for the AF3 dataset were lower than those for CF-T and CF-F. This result is consistent
280 with Section 3.3, which shows comparatively lower PCC for AF3. Among the scores
281 evaluated, pDockQ2 consistently exhibited the lowest AUC (Fig. 4B) and also
282 accuracy (Suppl. Table 2) across all three datasets, in agreement with the distributions
283 shown in Fig. 3. Considering the low correlation of pDockQ2 with DockQ (Fig. 3) its
284 threshold determination is less reliable. This is especially clear for the CF-F dataset,
285 where the threshold is below 0.1. On the other hand, the highest AUC (Fig. 4B) and
286 accuracy (Suppl. Table 2) values were observed for the ipTM and *model confidence*
287 scores, in agreement with the findings of Section 3.3.
288 To assess AF3 models, there were guidelines for the cutoffs of ipTM in the frequently
289 asked questions of the AlphaFold Server (Accessed October 29, 2024.
290 https://alphafoldserver.com/faq). ipTM > 0.8 indicates high-quality predictions, ipTM
291 < 0.6 suggests likely failed predictions, and ipTM in 0.6 - 0.8 range falls into a “gray
292 zone” where predictions could be correct or incorrect.
293 However, the importance of providing a specific cutoff to separate correct from
294 incorrect models arises when the model’s ipTM falls within the gray zone. In all three
295 examples shown in Fig. 4C, the ipTM scores fall within the gray zone defined in

296 AlphaFold server, yet the outcomes vary significantly. In Fig. 4C (i and iii), the
297 predictions by CF-T (target PDB: 7PPZ) and AF3 (target PDB: 6MM8) do not pass the
298 ipTM corresponding cutoffs (0.70 and 0.75, respectively) and are categorized as
299 'incorrect' based on DockQ scores. In contrast, Fig. 4C(ii) shows a CF-F prediction
300 (target PDB: 7BZK) for which the ipTM score surpasses our cutoff of 0.65, with the
301 DockQ score classifying this prediction as of 'high' quality. Therefore, for the subset of
302 structures of CF-T, CF-F and AF3 with ipTM values in the range of 0.6-0.8, we
303 evaluated the accuracy with cutoffs determined by ROC analyses (CF-T: 0.7; CF-F:
304 0.65; AF3: 0.75, Fig. 4B). This allowed us to differentiate between correct and incorrect
305 models even in the gray zone with accuracies of 0.81, 0.87 and 0.63 for CF-T, CF-F
306 and AF3, respectively (Suppl. Table 3).

A

Binary classification correct vs. incorrect			
Score	**CF-T**	**CF-F**	**AF3**
ipLDDT	80.0	80.0	85.0
iPAE	5.00	9.00	6.00
pTM	0.80	0.80	0.75
ipTM	0.70	0.65	0.75
mc	0.75	0.70	0.70
pDockQ2	0.35	0.05	0.45
VoroIF	0.65	0.65	0.60

B



C

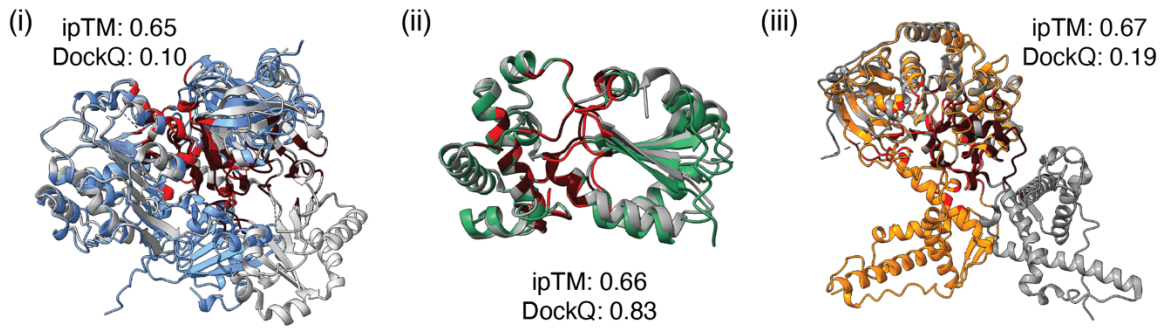


Fig. 4: ROC curves, determined cutoffs and example predictions. **A:** Determined thresholds for the scores distinguishing correct ('high', 'medium', 'acceptable' DockQ) and incorrect ('incorrect' DockQ) models for CF-T, CF-F and AF3. **B:** ROC-curves and AUC of the raw scores (ipLDDT, iPAE, pTM, ipTM, *model confidence*, pDockQ2 and VoroIF) for CF-T, CF-F and AF3. The binary classification is based on the DockQ classification correct ('high', 'medium', 'acceptable' DockQ) and incorrect ('incorrect' DockQ) **C(i):** CF-T prediction (blue, interface: red) and target structure (PDB: 7PPZ) (gray, interface: dark red). **C(ii):** CF-F prediction (green, interface: red) and target structure (PDB: 7BZK) (gray, interface: dark red).

316 **C(iii):** AF3 prediction (orange, interface: red) and target structure (PDB: 6MM8) (gray,
317 interface: dark red).

318 ### 3.5. Combined scoring and testing

319 In addition to evaluating the individual scores and their discriminatory power, a
320 combined score was developed to improve the assessment of predicted protein
321 complex models. Various approaches were tested to establish a combined score (See
322 Methods). Given that *model confidence* and pDockQ2 are computed by combining
323 other evaluated scores, these scores were excluded.

324 A linear regression-based model for the combination of all scores (ipLDDT, iPAE, pTM,
325 ipTM, and VoroIF) was consistently identified as the best-performing (based on PCC)
326 combined score across all three datasets (CF-T, CF-F and AF3), and referred to as
327 C2Qscore (Combined Complex Quality score). This score was selected for further
328 analyses (Suppl. Table 4, for weights, bias and maxima see Suppl. Tab. 5-7). Wilcoxon
329 signed-rank test indicated that C2Qscore provided a statistically significant
330 improvement compared to the unweighted combination of the individual scores (Suppl.
331 Table 8).

332 Finally, ROC curve analysis was conducted to determine a cutoff for differentiating
333 between correct models and incorrect ones (Suppl. Table 9). This analysis yielded a
334 cutoff of 0.56 (AUC: 0.95) for models predicted using ColabFold with templates, 0.52
335 (AUC: 0.96) for models predicted without templates, and 0.65 (AUC: 0.94) for models
336 predicted by AF3.

337 To further assess the performance of C2Qscore, it was tested on an independent
338 dataset of 40 recently released structures of heterodimeric protein complexes
339 determined by X-ray crystallography (resolution better than 3 Å). DockQ and all
340 prediction-based scores (including the combined ones) were calculated for 160

models for CF-T, 159 models for CF-F, and 185 models for AF3. The cutoffs defined in Sections 3.4 and 3.5 were applied to all scores to calculate the accuracy and precision (Fig. 5). C2Qscore demonstrated an improvement in model classification over the unweighted score (simple average of the individual scores) for CF-T (Δaccuracy: +0.01) and AF3 (Δaccuracy: +0.03) and individual scores (Supp. Fig. 6). For CF-F, the unweighted combined score and C2Qscore showed equal performance (Δaccuracy: 0.0), though C2Qscore provided a clearer separation between incorrect and correct models compared to the unweighted score.

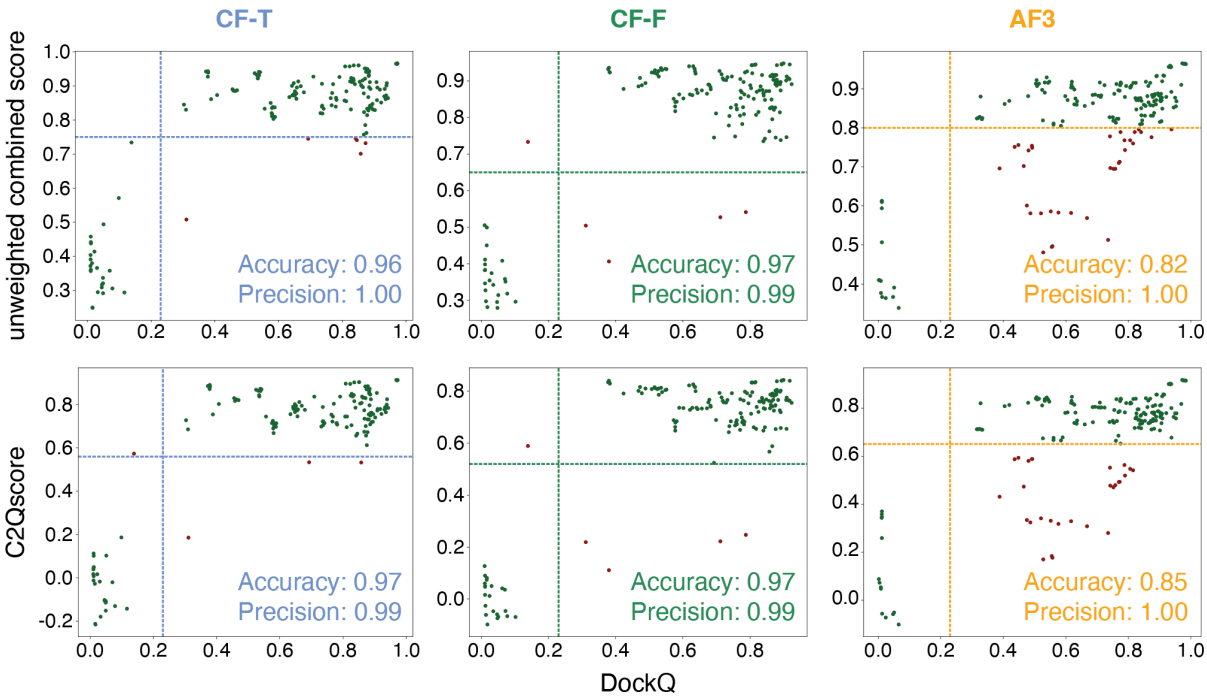


Fig. 5: Unweighted combined score and C2Qscore for the X-ray crystallography test set for CF-T, CF-F and AF3. True positives and true negatives are plotted in dark green and false positives and false negatives in dark red. The in Section 3.5 determined cutoffs are shown as dotted lines in the plots.

3.6. Evaluating the scores on dimers from larger assemblies

Next, we evaluated the performance of the unweighted combined score and C2Qscore on a dataset of dimers derived from assemblies of higher order (≥ trimeric). This evaluation aimed to assess the performance of the prediction methods and the scoring

358 approaches on dimers, without prior knowledge of the biological context. The cutoffs
359 determined in Section 3.5 were applied. This set included a total of 156 dimeric target
360 structures derived from cryoEM assemblies at resolutions higher than 2.3 Å.
361 Here, as before, the performance of C2Qscore was similar to the unweighted score,
362 showing a mild improvement (statistically insignificant) in model classification based
363 on DockQ only for CF-T (Δaccuracy: 0.01), while a mild decrease in performances of
364 CF-F (Δaccuracy: -0.02) and AF3 (Δaccuracy: -0.01) (Fig. 6, Supp. Fig. 7). The scoring
365 of these models had a higher percentage of false negatives across all datasets
366 compared to the X-ray test set (Fig. 5, Fig. 6A, Supp. Fig. 7).
367 To understand the general decrease in accuracy compared to the X-ray test set, we
368 further analyzed the cryoEM models by investigating the distribution of DockQ and
369 C2Qscore in relation to the number of possible configurations of the dimeric complex
370 derived from large assemblies. In some cases, the target chains of a dimer, which is
371 part of a larger complex, has homologous counterparts within the assembly forming
372 different configurations (Fig. 6B). Since AlphaFold2 and AlphaFold3 only accept the
373 sequence as input, one cannot know in advance which of the available configurations
374 correspond to the predicted structure. Fig. 6C depicts the dimeric complex formed by
375 the chains D and I of the large assembly with PDB-code 9ESY. Chain I is homologous
376 to chain A in the assembly, which also interacts with chain D, resulting in two possible
377 configurations for the target complex. The CF-T prediction closely resembles the dimer
378 formed by chains A and D (DockQ 0.74), but because DockQ is calculated for the
379 complex formed by chains I and D, DockQ is very low (0.01). In contrast, C2Qscore
380 for this prediction is relatively high, at 0.72.
381 In a typical prediction scenario, where no ground truth is available, the number of
382 potential configurations of a complex is unknown. Therefore, it is crucial to assess the

383 reliability of the scores when various spatial arrangements for two chains to form a
384 dimer ("configuration") exist. To address this, we used the "large assembly set" and
385 analyzed the DockQ and C2Qscore versus the number of configurations per target
386 dimer for the entire "large assembly" set. The dataset was filtered to include only those
387 templates for which more than one dimeric configuration exist, specifically those with
388 two or three configurations. Furthermore, the predictions were required to be classified
389 as correct based on DockQ (DockQ ≥ 0.23) for one of the existing configurations. It is
390 important to note that the presented DockQ scores (Fig. 6B) correspond to the
391 alignment between the target structure and the prediction. This, by definition, will result
392 in low DockQ values in cases where the prediction closely resembles one of the
393 alternative configurations but is compared to the target configuration.
394 For cases with two potential configurations of the dimer (n=55) the DockQ showed a
395 bimodal distribution, featuring very high and very low scores (Fig. 6B). However, the
396 C2Qscore distribution differed significantly, with many models scored with very high
397 or medium-to-high scores. For cases involving three possible configurations in CF-T
398 (n=33), the DockQ score peaked at low scores (due to more "different" configuration
399 from the target), while the distribution of the C2Qscore was spread in the entire range
400 of values. Similar trends were observed in the CF-F and AF3 dataset (Suppl. Fig. 8).
401 These findings suggest that both AF2 and AF3 struggle to differentiate between
402 probability and certainty when multiple configurations of a dimer are present. As a
403 result, lower scores (medium-to-high or lower) may be assigned to highly accurate
404 predictions, making the scores less reliable. This could explain the general decrease
405 in accuracy of the models in the large assembly set compared to the X-ray test set.

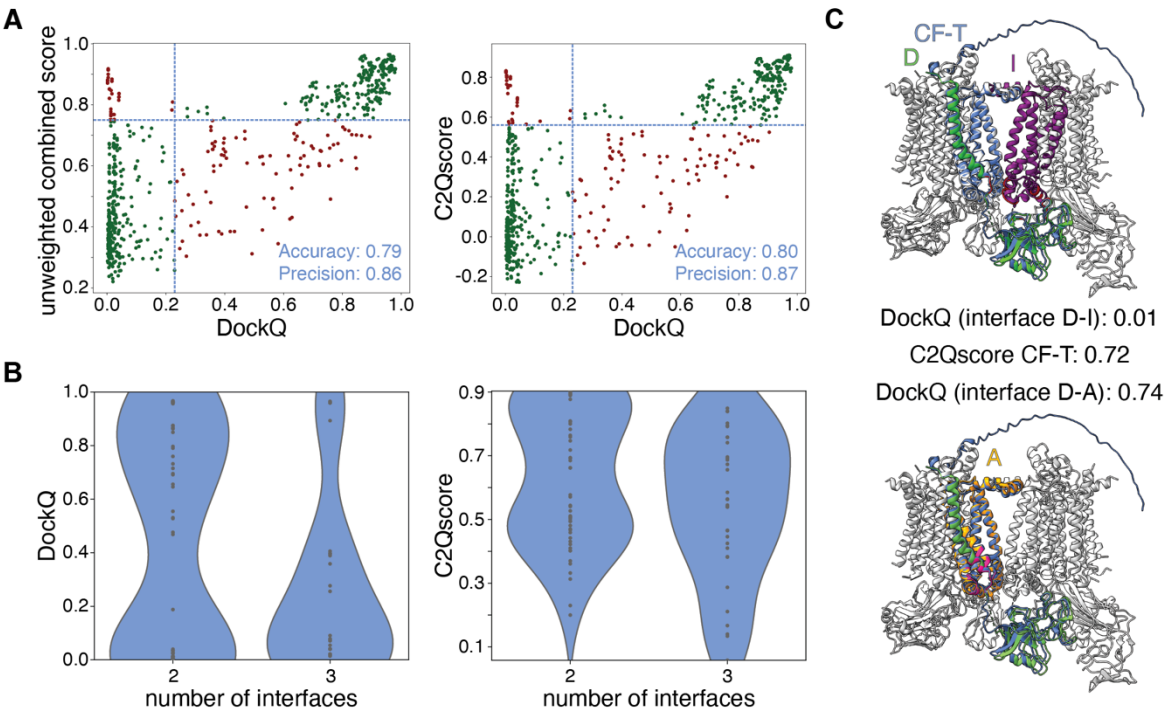


Fig. 6: A: Unweighted combined score and C2Qscore for the large assembly test set for CF-T. True positives and true negatives are plotted in dark green (upper right and lower left) and false positives and false negatives in dark red (upper left and lower right). The cutoffs we determined based on the benchmark set are shown in light blue. **B:** Distribution of DockQ and C2Qscore for CF-T cases of two (n=55) or three (n=33) potential configurations, for which the prediction resembles one of the configurations. **C:** CF-T prediction superposed on the large assembly (PDB: 9ES7; chains D&I). The prediction (blue), the target chains D (green), I (purple) and the homologue of chain I, chain A (gold), are highlighted on the assembly (gray). The interface of chain D and I (red, top, target) and chain D and A (hot pink, bottom, other possible configurations) are shown. The DockQ for the alignment of the prediction with the complex D-I and D-A and C2Qscore for the prediction are reported.

3.7. C2Qscore as a tool

The algorithm for calculating the C2Qscore for AlphaFold2/ColabFold and AlphaFold3 models has been integrated as a tool into our ChimeraX plug-in PICKLUSTER v2.0 (Pettersen et al. 2021; Genz et al. 2023). The new PICKLUSTER v2.0 identifies and clusters protein interfaces based on distance, providing a targeted approach for their

423 analysis. It offers access to a range of scoring metrics, including the following scores
424 analyzed in this work: pLDDT, PAE, ipLDDT, iPAE, pTM, ipTM, and *model confidence*.
425 Here, PICKLUSTER v2.0 has been updated to support predictions generated by
426 AlphaFold3, and the newly developed C2Qscore has been integrated as a feature in
427 the user-interface.

428 **4. Discussion**

429 In this study, we presented a comprehensive evaluation of commonly used scores for
430 assessing protein complex models and compared three prediction methods:
431 ColabFold with templates (CF-T), ColabFold without templates (CF-F), and
432 AlphaFold3 (AF3). Using a benchmark set of 325 heterodimeric, high-resolution target
433 structures from the PDB, we analyzed the ability of these methods to produce accurate
434 models based on the CAPRI DockQ classifications. We also assessed the
435 effectiveness of the following scores: ipLDDT, iPAE, pTM, ipTM, *model confidence*,
436 pDockQ2 and VoroIF, in distinguishing correct from incorrect predictions.

437 Overall, we found that ipTM and *model confidence* consistently show the highest PCC
438 with DockQ, highest AUC for classifying models as correct or incorrect and high
439 accuracies across all datasets (CF-T, CF-F, and AF3), suggesting these scores as the
440 most reliable ones to assess heterodimeric protein complex models with both
441 ColabFold and AlphaFold3.

442 Furthermore, in general the interface-specific scores, ipLDDT and iPAE, showed
443 higher PCCs with DockQ than their global counterparts, pLDDT and PAE, consistent
444 with the findings of (Yin et al. 2022). This highlights the importance of interface-specific
445 metrics for assessing protein complex models.

446 Interestingly, our results reveal that also iPAE exhibits a high PCC with DockQ, ranking
447 just below ipTM and *model confidence*. Additionally, we observe a strong correlation

448 between PAE and pTM, and between iPAE and ipTM. This can be attributed to the
449 method used to compute ipTM and pTM, both of which are derived from the PAE
450 matrix (Evans et al. 2022). However, during this process, the values within the PAE
451 matrix are weighted, such that residue pairs with lower PAE values contribute more
452 significantly to ipTM and pTM. When averaged across the entire structure, this
453 weighting effect is diminished.
454 Importantly , the cutoffs we define for differentiating correct from incorrect models
455 using ipTM fall into the gray zone thresholds proposed by the AlphaFold web server
456 (Accessed October 29, 2024, https://alphafoldserver.com/faq). However, while
457 AlphaFold's ipTM gray zone leaves model quality open to interpretation, our cutoffs
458 offer a practical guide, helping users to assess the reliability of protein complex models
459 based on the chosen structure prediction method and settings. In this study, we
460 establish a new, high-accuracy cutoff specifically for ColabFold, not only for ipTM but
461 for all scoring metrics, to better differentiate between correct and incorrect models.
462 Notably, the determination of cutoffs was also effective for AF3, though with lower
463 accuracy compared to CF-T and CF-F. Along with lower AUC, accuracy values and
464 lowest mean PCC between the individual scores and DockQ, these findings suggest
465 an overall weaker performance of assessment scores for AlphaFold3 compared
466 ColabFold. This may be due to the differences in the network architecture and
467 confidence estimation of AlphaFold3 compared to AlphaFold2-multimer (Abramson et
468 al. 2024).
469 According to our benchmark, CF-T and AF3 produce similarly reliable predictions. This
470 contrasts with the findings of (Abramson et al. 2024), which reported that AlphaFold3
471 improved protein complex accuracy compared to AlphaFold-multimer (v.2.3), the

472 version employed in ColabFold v.1.5.2 (which is also the version used in this work
473 (CF-T)).
474 Not surprisingly, our results show that as the prediction quality (DockQ) decreases,
475 the agreement among the prediction methods also declines. Interestingly, this
476 suggests that for cases of lower prediction quality, in particular when the scores are
477 around the “acceptable” zone, using multiple prediction methods and comparison of
478 the results may be beneficial.
479 We analyzed our benchmark for cases where the asymmetric unit (AU) differed from
480 the biological assembly (BA) in the PDB. Interestingly, in such instances, we observed
481 that all three methods could predict either the AU or the BA with high confidence, but
482 without prior knowledge, it would not be possible to determine which outcome reflects
483 the correct structure. This scenario becomes particularly problematic when the
484 predicted AU is correct based on scoring metrics but does not correspond to the
485 functional BA, potentially leading to false interpretations in subsequent experiments.
486 Such cases underline the importance of cautious interpretation of predictions,
487 particularly in scenarios where the biological relevance of the AU and BA is unclear
488 (Schweke et al. 2023).
489 To improve the ability of assessing protein complex models, we developed a combined
490 score, C2Qscore (Combined Complex Quality score) based on a linear regression
491 model that included ipLDDT, iPAE, pTM, ipTM and VoroIF. C2Qscore improved the
492 ability to assess model quality, particularly for CF-T and AF3 model predictions tested
493 on a X-ray crystallography dataset, where it outperformed the unweighted combination
494 of the individual scores.
495 Testing of C2Qscore on the ‘large assembly’ dataset revealed an increased
496 percentage of false negatives and a reduced classification accuracy across all

497 prediction methods compared to the X-ray test set. This suggests that when multiple
498 configurations of a dimer are possible, AlphaFold2 and AlphaFold3 may give lower
499 scores (medium-to-high rather than high) to highly accurate models, compromising the
500 reliability of the scores.

501 Finally, it is important to note that this study focused exclusively on protein pairs that
502 are known to interact, using the scoring metrics to evaluate the quality of the predicted
503 complexes. Other studies have applied these scores to determine whether a protein
504 pair interacts or not (Schweke et al. 2023). The latter approach should be used with
505 caution, as studies have shown that ipTM is sensitive to variations in sequence
506 constructs, leading to different ipTM scores despite unchanged predicted interface
507 contacts (Bret et al. 2024; Pándy-Szekeres et al. 2024). This suggests that while a
508 high ipTM score indicates confidence in the prediction, a low score does not
509 necessarily imply that protein pairs do not interact and scores in the grey zone may or
510 may not imply interaction. Tools like PPIscreenML (Mischley et al. 2024) may provide
511 potential solutions to address this challenge.

512 **5. Conclusion**

513 Our study shows both the strengths and limitations of individual commonly used
514 assessment scores for evaluating heterodimeric protein complex models produced by
515 ColabFold with and without templates and by AlphaFold3. Our PICKLUSTER v2.0
516 plug-in enables the usage of these scores in an interactive and easy manner. Overall,
517 we recommend ipTM and *model confidence* for general use due to their high
518 correlation with DockQ. However, the interpretation of the scores needs to be
519 performed carefully, particularly for AlphaFold3 models. Notably, our analysis was
520 conducted exclusively on heterodimeric protein complex models without the
521 consideration of multiple interfaces. Further research should investigate the

522 transferability of the results to higher-order complexes. Moreover, exploring different
523 interfaces could reveal how scores perform in situations where interface regions differ
524 significantly in quality. With AlphaFold3 also enabling the prediction of protein-ligand,
525 protein-lipid, and protein-nucleotide interactions, future research could focus on
526 scoring metrics for these types of complexes. This also includes complexes involving
527 antibodies, which were excluded from the analysis in this work. Refining scoring
528 methods and expanding their applicability to diverse complex types and interface
529 characteristics will be for computational protein complex modeling to reach its full
530 potential.

531 **6. Methods**

532 **6.1. Generation of the benchmark dataset (Suppl. Fig. 1)**

533 PDB entries were initially filtered to include hetero-dimeric complexes with a release
534 date after 30.04.2018 (last date used in the AlphaFold training set), a release data
535 before 06.11.2023, and a refinement resolution ≤3.0 Å. The entries were grouped by
536 a sequence identity of >30% using PDB default settings (MMseqs2 (Steinegger and
537 Söding 2017)). The asymmetric unit of representative structures from groups that
538 satisfied the specified criteria were downloaded from the PDB, yielding a total of 1,381
539 complexes.

540 In the subsequent step, interacting complexes were defined as two chains having at
541 least three residues from each chain interacting with residues of the other chain, with
542 heavy atom distance cutoff of 5 Å, using PICKLUSTER (Genz et al. 2023). Non-
543 interacting complexes, as well as any chain with less than 50 residues, were filtered
544 out using the PICKLUSTER pipeline. This process resulted in a total of 580 protein
545 complexes.

546 In the next step, a PDB search and a manual analysis were conducted in order to filter
547 out entries containing antibodies or nanobodies. Subsequently, to ensure that the
548 dimers could be classified as heterodimeric complexes, homology within the
549 complexes was assessed by application of CD-HIT (Fu et al. 2012) with a 30%
550 sequence identity threshold, resulting in 397 complexes. Moreover, only templates for
551 which one BA assigned in the PDB equaled the AU were included in the benchmark
552 set, resulting in 325 target complex structures in the final benchmark set.

553 **X-ray crystallography test set**

554 The X-ray crystallography test set (X-ray test set) was curated following the same
555 pipeline as for the benchmark set, with only slight adjustments to the selection criteria
556 (Suppl. Fig. 1). Target structures present in the benchmark set were excluded from
557 the X-ray test set, resulting in 40 target complex structures in the X-ray test set.

558 **Large assembly test set**

559 The set of dimers from larger assemblies ("large assembly test set") was curated
560 following the protocol established for the benchmark set (Suppl. Fig. 1), with the
561 following modifications to the selection criteria: the dimers were derived from higher-
562 order assemblies (≥ trimeric) with a resolution ≤2.3 Å resolved by cryoEM. A stricter
563 resolution cutoff was implemented compared to the original benchmark set due to the
564 potential variability in local resolution relative to the reported global resolution in
565 structures resolved by cryoEM (Vilas et al. 2020). To ensure the non-redundancy of
566 dimers derived from the same larger assembly, the homology of the dimers was
567 evaluated using CD-HIT (Fu et al. 2012) with a 30% sequence identity threshold,
568 resulting in 156 dimeric target structures included in the large assembly test set.

569

570

571 **6.2. Model generation**

572 Three AI-based structure prediction approaches were used: CF-F, CF-T and AF3, to
573 create three corresponding datasets.

574 For the CF-F dataset structural models were predicted from sequence using local
575 ColabFold v1.5.2 (Mirdita et al. 2022). Default parameters were applied for multiple
576 sequence alignment (MSA) generation (MSA mode: mmseqs_uniref_env) and three
577 recycles were performed. For each template, five models were generated, all of which
578 were included in the dataset. Relaxation was performed to all five models.

579 For the CF-T dataset structural models were predicted using the same criteria as for
580 the CF-F dataset. The only difference was the activation of the template option.

581 For the AF3 dataset structural models were predicted using the AlphaFold3 web server
582 (https://alphafoldserver.com/) (Abramson et al. 2024) with default settings, which
583 includes the usage of templates.

584 **6.3. Model preparation**

585 For the assessment of the model quality with DockQ, after prediction with ColabFold
586 or AlphaFold3 residues were renumbered according to the template residue
587 numbering, utilizing the fix_numbering.pl script provided in the DockQ package. An in-
588 house Python script was employed to rename chain identifiers in the models, ensuring
589 alignment between the template and the model chains.

590 AF3 models were initially predicted in mmCIF format, and to enable for the assessment
591 with DockQ, they were converted into PDB format with a Python script that uses the
592 *Biopython (Cock et al. 2009)* package.

593 **6.4. Reference-based score**

594 DockQ v1.0 (Basu and Wallner 2016) (https://github.com/bjornwallner/DockQ)
595 (https://github.com/bjornwallner/DockQ) was used to evaluate the quality of the

596 predicted models, by comparing the predicted models of a complex against the native
597 complex structure. The score combines the criteria from the CAPRI competition
598 (Lensink and Wodak 2013) into a single value. Scores range from 0 to 1, with 1
599 indicating a perfect match between the interface of the predicted model and that of the
600 native complex structure. Threshold values of 0.23, 0.49, and 0.8 were used to
601 categorize predictions as acceptable, medium, and high quality respectively,
602 according to CAPRI standards. DockQ scores were determined by comparing the
603 predicted models of a complex against the native complex structure. DockQ v1.0 was
604 installed according to https://github.com/bjornwallner/DockQ.

605 **6.5. Prediction-based scores**

606 The scores used for assessment of models are: pLDDT (predicted local difference
607 distance test) (Jumper et al. 2021), PAE (predicted alignment error) (Evans et al.
608 2022), ipLDDT (interface pLDDT) (Jumper et al. 2021; Genz et al. 2023), iPAE
609 (interface PAE) (Evans et al. 2022; Genz et al. 2023), pTM (predicted TM-score)
610 (Jumper et al. 2021), ipTM (interface pTM) (Evans et al. 2022), *model confidence*
611 (Evans et al. 2022), pDockQ2 (Zhu et al. 2023) and VoroIF (VoroIF-GNN) (Olechnovič
612 and Venclovas 2023 Jul 21). The details about each score, including the range, what
613 it measures and whether it is interface specific or not, are provided in Supp. Table. 1.

614 **6.6. Performance metrics**

615 To measure the overall performance of scores over the different datasets we used a
616 number of approaches:

617 *Pearson correlation coefficient:* The Pearson correlation coefficient (PCC) gives an
618 indication of the strength of the linear relationship between two datasets. The PCC
619 ranges from -1 to +1 with 0 indicating no correlation and correlations of -1/+1 implying
620 an exact linear relationship. The PCC only detects linear dependencies between the

621 datasets. Positive values indicate a correlation whilst negative values imply anti-
622 correlation (Benesty et al. 2009).
623 The PCC was calculated using the stats package provided in the *SciPy* v1.11.3 Python
624 package (Virtanen et al. 2020).
625
626 *ROC-curves and AUC:* To identify the best performing cutoff for the scores, a Receiver
627 Operator Curve (ROC) analysis was conducted. The optimal cutoff for the score was
628 determined by maximizing the separation between correctly and incorrectly assessed
629 models, calculated through the True Positive Rate (TPR) and False Positive Rate
630 (FPR), which were plotted as the ROC curve. The best performing cutoff was identified
631 as the point with the maximum distance from the diagonal in the ROC curve (Melo and
632 Sali 2007). For each point on the ROC curve, the cutoff values were incrementally
633 increased depending on the score type (pLDDT/ipLDDT: 5; iPAE/PAE: 1; all other
634 scores: 0.05). The Area Under the Curve (AUC) was calculated to quantify the
635 accuracy of the metric being evaluated, where an AUC of 0.5 indicates random
636 performance, and an AUC of 1 represents perfect discrimination. Both the ROC curves
637 and AUCs were computed using a Python script utilizing the metrics module from the
638 *Scikit-learn* v1.3.0 Python package (Pedregosa et al. 2011).
639
640 *Accuracy:* Accuracy is defined as the ratio of (TP+TN)/(TP+TN+FP+FN), where TP
641 represents the number of true positives, TN the number of true negatives, FP the
642 number of false positives and FN denotes the number of false negatives. The metric
643 is used to evaluate the performance of a model, indicating how frequently the model
644 makes correct predictions. It is determined by comparing the number of correct
645 predictions to the total number of predictions made by the model. The metric is

646 expressed as a value between 0 and 1, with 1 indicating the optimal outcome. The
647 accuracy provides an overall measure of the reliability of a model (Powers 2020).
648
649 *Precision:* Precision is defined as the ratio TP/(TP+FP). Precision quantifies the
650 proportion of correctly identified positive samples out of all samples predicted as
651 positive. It therefore reflects the ability of the classifier to avoid falsely classifying
652 negative samples as positive. A precision value of 1 indicates perfect performance,
653 with no false positives, while a value of 0 represents the worst outcome, with all
654 predicted positives being incorrect. Precision shows how often a model is correct when
655 predicting the target class (Powers 2020).
656
657 *Wilcoxon signed rank (WSR) test:* The WSR test is a nonparametric test procedure for
658 the evaluation of the significance of paired data (Woolson 2005). This non-parametric
659 test was chosen due to its ability to compare matched data without assuming
660 normality. The p-value was calculated to determine the significance of the observed
661 difference. A significance threshold of $p<0.05$ was applied. The WSR test was
662 implemented using the stats package provided in the *SciPy* Python package (Virtanen
663 et al. 2020).

664 **6.7. Influence of the distance threshold for the definition of the interface on**
665 **the scores**

666 To determine the interface PAE (iPAE) and interface pLDDT (ipLDDT), the interface
667 was defined as residues with less than 5Å (Evans et al. 2022) distance between heavy
668 atoms (non-hydrogen atoms) of the two chains in the heterodimeric structure, which
669 was calculated with PICKLUSTER (Genz et al. 2023).

670 To investigate the impact of the distance parameter on ipLDDT and iPAE, we chose
671 to do this analysis on the dataset generated using CF-T, which had 1,612 models in
672 total. We then computed ipLDDT and iPAE on these models using distance thresholds
673 in a range of 5 to 9Å (1Å steps).

674 **6.8. Linear and Lasso regression**

675 Linear regression was tested as the first approach for a weighted combined score. It
676 fits a linear model with coefficients by minimizing the residual sum of squares between
677 the target values in the datasets and the ones predicted by the approximation of the
678 linear model. A Lasso regression model was evaluated as the second approach to
679 define a weighted combined score. This approach uses a linear model that is fitted
680 under the assumption that only few of the coefficients are non-zero. Therefore, it
681 identifies the most relevant features and reduces the number of input variables.
682 Linear and Lasso regression were performed by using the linear_model package
683 provided in the *Scikit-learn* Python package (Pedregosa et al. 2011).

684 **6.9. Combined scoring**

685 To enable a combined score, both ipLDDT and iPAE were normalized between 0 and
686 1, representing the optimal outcome. Subsequently, various approaches were tested
687 to establish a combined score demonstrating superior model assessment capacity.
688 Given that *model confidence* and pDockQ2 are computed by combining other
689 evaluated scores, these scores were excluded.
690 In the first approach we selected the best combination of scores (ipLDDT, iPAE, pTM,
691 ipTM, and VoroIF) based on PCC with DockQ as the evaluation criterion. All possible
692 combinations without repetition of equally weighted scores within each set were
693 tested, and PCCs were computed for each. In the second approach we employed
694 linear regression, applied to both the full combination of scores and the best-

performing combination identified in the first approach. Finally, Lasso regression was tested as an approach for improvement over using individual scores.

7. Acknowledgments and Funding

We thank the Topf group, especially Thomas Mulvaney, Karen Manalastas-Cantos and Natan Nagar, for discussions and valuable feedback. This work was supported by the cooperation of Leibniz Institute of Virology Strategic Incentive Program and the Leibniz ScienceCampus InterACt (funded by the BWFGB Hamburg and the Leibniz Association), the BMBF Computational Life Sciences project (ASPIRE, 031L028), the Landesforschungsförderung Hamburg (HamburgX) and DFG CRC1648.

Conflict of interest: none declared.

8. References

Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, et al. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 630(8016):493–500.

Baker D, Sali A. 2001. Protein structure prediction and structural genomics. Science. 294(5540):93–96.

Basu S, Wallner B. 2016. DockQ: A Quality Measure for Protein-Protein Docking Models. PLoS One. 11(8):e0161879.

Benesty J, Chen J, Huang Y, Cohen I. 2009. Pearson Correlation Coefficient. In: Noise Reduction in Speech Processing. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 1–4.

Bret H, Gao J, Zea DJ, Andreani J, Guérois R. 2024. From interaction networks to interfaces, scanning intrinsically disordered regions using AlphaFold2. Nat Commun. 15. doi:10.1038/s41467-023-44288-7. https://www.nature.com/articles/s41467-023-44288-7.

Bryant P, Pozzati G, Elofsson A. 2022. Improved prediction of protein-protein interactions using AlphaFold2. Nat Commun. 13(1):1265.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 25(11):1422–1423.

Cozzetto D, Kryshtafovych A, Ceriani M, Tramontano A. 2007. Assessment of predictions in the model quality assessment category. Proteins. 69 Suppl 8(S8):175–183.

Edmunds NS, Alharbi SMA, Genc AG, Adiyaman R, McGuffin LJ. 2023 Jun 14. Estimation of model accuracy in CASP15 using the ModFOLDdock server. Proteins.

doi:10.1002/prot.26532. http://dx.doi.org/10.1002/prot.26532.

Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, Žídek A, Bates R, Blackwell S, Yim J, et al. 2022. Protein complex prediction with AlphaFold-Multimer. bioRxiv.:2021.10.04.463034. doi:10.1101/2021.10.04.463034. [accessed 2022 Sep 5]. https://www.biorxiv.org/content/biorxiv/early/2022/03/10/2021.10.04.463034.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 28(23):3150–3152.

Genz LR, Mulvaney T, Nair S, Topf M. 2023. PICKLUSTER: a protein-interface clustering and analysis plug-in for UCSF ChimeraX. Bioinformatics. 39(11):btad629.

Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. 2006. Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins. 65(3):712–725.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. Nature. 596(7873):583–589.

Lensink MF, Wodak SJ. 2013. Docking, scoring, and affinity prediction in CAPRI. Proteins. 81(12):2082–2095.

Mariani V, Biasini M, Barbato A, Schwede T. 2013. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 29(21):2722–2728.

Melo F, Sali A. 2007. Fold assessment for comparative protein structure modeling. Protein Sci. 16(11):2412–2426.

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. ColabFold: making protein folding accessible to all. Nat Methods. 19(6):679–682.

Mischley V, Maier J, Chen J, Karanicolas J. 2024. PPIscreenML: Structure-based screening for protein-protein interactions using AlphaFold. doi:10.7554/elife.98179.1. http://dx.doi.org/10.7554/elife.98179.1.

Namba K, Makino F. 2022. Recent progress and future perspective of electron cryomicroscopy for structural life sciences. Microscopy (Oxf). 71(Supplement_1):i3–i14.

Olechnovič K, Venclovas Č. 2019. VoroMQA web server for assessing three-dimensional structures of proteins and protein complexes. Nucleic Acids Res. 47(W1):W437–W442.

Olechnovič K, Venclovas Č. 2023 Jul 21. VoroIF-GNN: Voronoi tessellation-derived protein-protein interface assessment using a graph neural network. Proteins. doi:10.1002/prot.26554. http://dx.doi.org/10.1002/prot.26554.

Pándy-Szekeres G, Taracena Herrera LP, Caroli J, Kermani AA, Kulkarni Y, Keserű GM, Gloriam DE. 2024. GproteinDb in 2024: new G protein-GPCR couplings, AlphaFold2-multimer models and interface interactions. Nucleic Acids Res. 52(D1):D466–D475.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Louppe G, Prettenhofer P, Weiss R, et al. 2011. Scikit-learn: Machine Learning in Python. J Mach Learn Res. abs/1201.0490:2825–2830.

Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE.

2021. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci. 30(1):70–82.

Pierce BG, Hourai Y, Weng Z. 2011. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. PLoS One. 6(9):e24657.

Powers DMW. 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv [csLG]. http://arxiv.org/abs/2010.16061.

Schulz GE, Schirmer RH. 2013. Principles Protein Structure. Berlin/Heidelberg, Germany: Springer Science & Business Media.

Schweke H, Xu Q, Tauriello G, Pantolini L, Schwede T, Cazals F, Lhéritier A, Fernandez-Recio J, Rodríguez-Lumbreras LA, Schueler-Furman O, et al. 2023. Discriminating physiological from non-physiological interfaces in structures of protein complexes: A community-wide study. Proteomics. 23(17):e2200323.

Smyth MS, Martin JH. 2000. X ray crystallography. Mol Pathol. 53(1):8–14.

Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 35(11):1026–1028.

Studer G, Tauriello G, Schwede T. 2023 Oct 19. Assessment of the assessment-All about complexes. Proteins. doi:10.1002/prot.26612. http://dx.doi.org/10.1002/prot.26612.

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, et al. 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 50(D1):D439–D444.

Vilas JL, Heymann JB, Tagare HD, Ramirez-Aportela E, Carazo JM, Sorzano C. 2020. Local resolution estimates of cryoEM reconstructions. Curr Opin Struct Biol. 64:74–78.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 17(3):261–272.

Woolson RF. 2005. Wilcoxon Signed-Rank Test. In: Encyclopedia of Biostatistics. Chichester, UK: John Wiley & Sons, Ltd. https://onlinelibrary.wiley.com/doi/abs/10.1002/0470011815.b2a15177.

Xue LC, Dobbs D, Bonvin AMJJ, Honavar V. 2015. Computational prediction of protein interfaces: A review of data driven methods. FEBS Lett. 589(23):3516–3526.

Yin R, Feng BY, Varshney A, Pierce BG. 2022. Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. Protein Sci. 31(8):e4379.

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

Zhu W, Shenoy A, Kundrotas P, Elofsson A. 2023. Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. Bioinformatics. 39(7):btad424.