CASP16 protein monomer structure prediction assessment

Rongqing Yuan[1,2,3,*], Jing Zhang[1,2,*], Andriy Kryshtafovych[4], R. Dustin Schaeffer[2], Jian Zhou[3], Qian Cong[1,2,5,#], Nick V. Grishin[2,6,#]

[1]Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, USA.
[2]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, USA.
[3]Lyda Hill Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX, USA.
[4]Genome Center, University of California, Davis, California, USA.
[5]Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX, USA.
[6]Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX, USA.
[#]: Correspondence
[*]: These authors contributed equally.

Note: missing citations should be for papers appearing in the same CASP16 issue, which we will add in the paper production stage.

Abstract

The assessment of monomer targets in the Critical Assessment of Structure Prediction Round 16 (CASP16) underscores that the problem of single-domain protein fold prediction is nearly solved—no target folds were missed across all Evaluation Units. However, challenges remain in accurately modeling truncated sequences, irregular secondary structures, and interchain-induced conformational changes. The release of AlphaFold3 (AF3) during CASP16, and its effective integration by many groups, demonstrated its superiority over AlphaFold2 (AF2), particularly in confidence estimation and model selection. Additional improvements in multiple sequence alignments (MSAs) and construct design, i.e., selecting the optimal fragment of the full sequence for modeling, also contributed to enhanced prediction accuracy. The top three groups—all from the Yang lab—consistently outperformed others across CASP16 monomer targets, reflecting their robust modeling pipelines and successful adoption of AF3. CASP16 also introduced three new challenges: Phase 0, in which stoichiometry was withheld; Phase 2, which supplied ~8,000 MassiveFold models per target to test model selection strategies; and Model 6, which limited predictors to using MSAs provided by the organizers. While we evaluated group performance in these additional challenges, the insights gained were limited due to low participation and design flaws in the experiments. We suggest improvements for the organization of these challenges and encourage broader engagement from the prediction community. The progress in monomer modeling from CASP15 to CASP16 was very subtle, but more groups in CASP16 were able to outperform ColabFold, reflecting increased expertise in optimizing AF2 and the growing adoption of AF3. We anticipate that the recent release of the AF3 source code will stimulate future progress through user-driven optimization and innovations in model architecture. Finally, model ranking remains a persistent weakness across most groups, highlighting a critical area for future development.

Keywords

CASP16, protein structure prediction, monomer, AlphaFold2, AlphaFold3, quality estimation, stoichiometry, structure sampling

Introduction

Starting from 1994, the Critical Assessment of Structure Prediction (CASP) experiments[1], conducted biannually, have played a pivotal role in establishing a community-wide standard for protein structure prediction. CASP evaluates the performance of emerging computational methods, identifies the promising trends, and pushes the community's progress forward. The design of CASP experiments is as follows. Before the protein structures—determined through X-ray crystallography, cryo-electron microscopy, or nuclear magnetic resonance (NMR)—are publicly released and deposited into the Protein Data Bank (PDB)[2,3], their information is submitted to the CASP organizers. The organizers release suitable sequences for prediction. Participating groups submit predicted protein 3D structures based on the sequences. Subsequently, independent assessors are invited to evaluate the performance blindly using standardized metrics and statistical models, ranking groups based on their accuracy. More importantly, the assessors will also highlight significant methodological advances and pinpoint areas in need of further development[4].

The continuous progress in this field, in particular, the introduction of coevolution-based contact prediction and deep learning methods, has culminated in the breakthrough in protein structure prediction by AlphaFold2 (AF2) during CASP14[5–9], which has stimulated numerous scientific discoveries and biomedical applications[10–22]. AF2 has become an essential component for subsequent method development among CASP predictors, and winning strategies in CASP15 all built on top of AF2[23–25]. At the beginning of the current CASP (CASP16), DeepMind released the AlphaFold3 (AF3) web server[26], and multiple CASP16 predictors thus incorporated AF3 in their structure prediction pipeline. Assessing whether AF3 represents another significant improvement in the field is therefore a critical assessment task of CASP16.

This assessment focuses on monomer structure prediction, which has been at the core of the CASP experiment since its establishment[1]. However, since AF2 nearly solved the single-chain protein structure prediction problem in 2020, modeling monomer proteins has become relatively less challenging in recent years. Consequently, many monomer targets in this CASP are derived from homo-oligomeric protein complexes, multi-protein complexes, and protein-nucleotide complexes. Good performance in these monomer targets requires accurate modeling of the protein complexes, and thus, evaluation of monomer prediction accuracy also reflects CASP predictors' ability in complex structure modeling.

To stimulate progress in the field, the following new challenges regarding protein and protein complex structure prediction were introduced in the current CASP:

1. Phase 0. In the traditional CASP routine (Phase 1 of CASP16), predictors are required to predict the 3D structure based on sequences and stoichiometry of a protein complex. CASP16 introduced Phase 0, where predictors were not provided with stoichiometry

information. Predictors thus needed to solve the additional challenge of stoichiometry prediction, a factor significantly influencing the oligomeric structure prediction accuracy [cite: our oligomer evaluation paper], thereby affecting the monomer targets derived from protein complexes.

2. Phase 2. CASP16 also introduced Phase 2, in which predictors were provided with a large number (in most cases, 8,040 per target) of predicted structures by MassiveFold[27]. Massive sampling approaches using AlphaFold-like methods, exemplified by AFsample from the Wallner group, were a successful strategy in CASP15[28,29]. Thus, providing every predictor with large pools of predicted models is expected to stimulate progress in model quality estimation and model selection.

3. Model 6. Another winning strategy in CASP15 was improving upon the default multiple sequence alignments (MSAs) of AF2[30,23,31]. To examine the impact of deep learning network architecture independent of MSA quality, CASP16 introduced a special "model 6" aside from the traditional models 1–5. For model 6, predictors were asked to build structures using MSAs generated by ColabFold[32]. This additional challenge minimizes the influence of MSA quality and allows a focused evaluation of architectural innovations.

We herein report our assessment of the monomer prediction category in CASP16. In the following sections, we first analyze the strategies used by different groups and find that using AF3, enhancing MSAs, and refining the "constructs" used for modeling were the winning strategies of CASP16. We then outline the features of the targets that influence the performance, highlight difficult targets, and infer potential causes for that. We provide the rankings of participants. Finally, we perform the analysis of results in the newly introduced CASP16 challenges. The results again demonstrate that protein monomer structure prediction is a largely solved problem[33], and improvement in the monomer prediction accuracy in CASP16 compared to CASP15 is only subtle. We identify aspects that are still challenging to the field, such as predicting structures for truncated sequences, modeling irregular secondary structures, and selecting the best models. These areas represent directions for further advances by the community.

Results and Discussions

Overall view of targets and prediction strategies

For monomer categories, 125 human expert groups and 36 server groups submitted predictions for 16, 42, and 27 targets from Phase 0, 1, and 2, respectively. The Phase 0 monomer targets and models were all derived from Phase 0 oligomer targets and predictions submitted for these targets. These targets were defined from 11 monomeric proteins, 16 homo oligomeric protein complexes, 8 hetero oligomeric protein complexes, and 5 protein-nucleotide (hybrid) complexes. These proteins were first parsed into domains using DomainParser[34], DDomain[35], SWORD[36], and Chainsaw[37]. Evaluation Units (EUs) were defined as single domains or joint domains if their relative orientations can be correctly predicted by most groups[38,39]. We defined 25, 74, and 46 EUs for Phase 0, 1, and 2 targets, respectively.

An overview of the EUs, groups, and each group's performance (measured by GDT_HA of the best model) is shown in **Fig. 1**. CASP encourages prediction of all targets to increase the

statistical power of performance comparisons between groups, and the established evaluation routines penalize for missed targets. We evaluated the performance of all groups, but only those that submitted on more than a certain percentage of all EUs will be presented in the figures and tables, and the threshold will be denoted in the context. In **Fig. 1**, we only show groups that have predicted more than 25% (19) Phase 1 EUs.

For most EUs, nearly all groups could generate high-quality models, and the performance differences are minimal. Some EUs turned out to be challenging for most groups (with a 95% quantile of GDT_TS for the best model of every group below 60, see **Fig. 3A**), but even in those cases, folds were accurately predicted by the best model (supplemental **Fig. S1**), again suggesting that the folding problem of single proteins is nearly solved.

Similarly to CASP15, there is no single group that stands out[30]. Each group can only generate close-to-best models (GDT_HA difference from the best model < 0.5) in 20% of all EUs (supplemental **Fig. S2**). Many leading groups utilized AF2 and AF3[6,26] as well as customized MSA inputs to improve the performance of these networks. In CASP16, no group used pure protein language model (PLM)-based methods, in contrast to CASP15, where two groups used them[30,40].

A total of 69 groups participating in the monomer structure prediction category submitted an abstract briefly explaining the methods they used in CASP16, allowing us to annotate the strategies used by each group (on top of **Fig. 1**). We partitioned groups based on the types of modeling approaches and compared the performance between different strategies. Majority groups used AF2 and nearly half used AF3 (**Fig. 2A**). Our analysis indicates that using either AF2 or AF3 yields a better performance than using neither of them, and that only using AF3 or a combination of AF2 and AF3 is better than only using AF2 (**Fig. 2B**). This suggests that AF2/3 remains the state-of-the-art in protein structure modeling and AF3 does show noticeable advantage over AF2. It is worth noting that during the CASP16 prediction period, AF3 was only released as a web server and CASP16 predictors could only control the input sequences (full proteins or segments), test different stoichiometry, and sample more models by registering more accounts (each account can submit 20 jobs per day) and using multiple random seeds. With the recent public release of AF3's source code and trained weights, we expect further improvement in this field as predictors adopt AF3 in their pipelines and explore better ways of utilizing it.

Furthermore, using customized MSAs also correlates with better prediction accuracy (**Fig. 2C**). Although such observed performance differences are incremental, the difference is statistically significant. These findings support that leveraging the power of AF2 and AF3, as well as better MSA, can lead to superior performance in monomer structure prediction.

By contrast, we do not observe a significant difference in performance between groups that used PLMs and those that did not. In fact, groups not using PLMs showed slightly better performance in terms of mean GDT_HA scores (**Fig. 2D**). However, we believe this should not be interpreted as a negative contribution of PLMs; rather, it indicates that better performance can be achieved by optimizing AF modeling strategies (e.g., input sequences, MSAs, broader sampling) without the need for PLMs. Many outstanding groups did not use PLM in CASP16, which aligns with the observation in CASP15 that only using PLM instead of AlphaFold-like networks usually leads to suboptimal performance. In CASP16, PLMs are mostly used to conduct homology search or provide embeddings for AlphaFold-like networks. The optimal role of PLMs in protein structure prediction remains an open and promising area for future exploration.

EU features affecting performance.

We analyzed structural and biological features of EUs that may affect the predictors' ability to predict structures. We first found that prediction accuracy is positively (but not linearly) correlated with MSA depth measured by *Neff* [41] (**Fig. 2E**). Deeper MSAs provide rich coevolutionary signals and increase the likelihood of obtaining high-quality models. However, deep MSAs are not always required for high-quality predicted structures. It is observed that AF2/3 can predict 3D structures of a small fraction of natural protein sequences to high quality without MSAs, indicating that AF2/3 also learned additional physicochemical rules about protein folding[42], which might be sufficient to fold a subset of proteins.

Another related factor associated with prediction quality is the phylogenetic origin of a target, which can be broadly categorized into prokaryotic, eukaryotic, and viral proteins (**Fig. 2F**). Generally, prokaryotic proteins yield the deepest MSAs, as a result of broader taxonomy diversity and faster evolution rates compared to eukaryotes[43–46]. Predictors can also effectively mine extensive metagenomic databases to obtain more sequences in addition to sequenced genomes[47,48]. In contrast, eukaryotic and viral proteins tend to have shallower MSAs. Additionally, the rapid evolution of some viral proteins[49] and their tendency to form obligate oligomers[50] also make them more challenging targets. Reflecting these expected challenges in generating high-quality models of viral proteins[51], they were not included in AlphaFold protein structure DataBase (AFDB)[52].

The experimental method used to determine protein structures also appears to influence prediction accuracy, and the performance on structures solved by X-ray crystallography is generally better. However, the apparent difference between X-ray and cryo-EM could be attributed to the inherent difficulty of the targets—those resolved by cryo-EM are often larger and more complex, which typically leads to lower prediction accuracy. There is only one NMR target in this CASP (T1226-D1, from turkey viral polyprotein), and the performance for that target is particularly poor (**Fig. 2G**). A close inspection of this target reveals that a flexible loop in T1226-D1 (supplemental **Fig. S3A**) is frequently modeled as an ordered alpha-helix. Because structures solved by NMR tend to show higher flexibility compared to those determined by other techniques[53], we think that the predicted structure might represent a true alternative conformation, a hypothesis that remains to be tested by additional experimental data.

We noticed a couple of targets, such as T1220s1-D1 and T1226-D1 (supplemental **Fig. S3A**) showing low agreement with predicted structures might be poorly resolved and show poor MolProbity scores (>3, and high values indicating more violations to physicochemical rules)[54]. We also found a positive correlation between lower MolProbity scores of targets and poorer agreement with predicted structures (**Fig. 2H**). This is in agreement with another observation that predictors' performance is better on structures having higher experimental resolution (**Fig. S3B**). As the performance of structure prediction tools improves, the discrepancy between experimental and predicted structures might be used to detect errors and help to resolve poorly resolved regions in experimental structures.

Finally, we found that targets derived from hetero oligomers and hybrid complexes are generally harder to predict than those from monomers and homo oligomers (**Fig. 2I**). Proteins extracted from complex structures frequently undergo conformational changes upon

intermolecular interactions and accurate modeling of their structures need to happen in the context of their binding partners. While structure prediction for single-chain proteins is a largely solved problem, 3D structures of many protein complexes remain challenging to predict in the post-AlphaFold era [cite: our oligomer evaluation paper], making monomeric targets derived from these complexes more difficult.

Potential challenges for current methods

We estimated the field's performance on each EU by the 95% quantile of GDT_TS among its models, denoted as GDT_TS95. We found that larger EUs tend to show lower GDT_TS95 (**Fig. 3A**). This is primarily because of the nature of GDT_TS calculated on a global scale without normalizing over protein length. Thus, the errors in the 3D space can propagate as the number of residues increases[55]. Taking this EU-size bias into consideration, we identified challenging EUs for most CASP16 predictors as outliers in the GDT_TS95 versus EU size plot (labeled dots in **Fig. 3A**). We manually studied these EUs to identify the weakness of existing methods and suggested future directions to improve structure predictions.

Potential conformational changes upon truncation: AF2 have been found to be insensitive to mutations and truncations[56,57], and target T1207-D1 might represent such a case. The T1207-D1 sequence originates from an Avian viral polyprotein (**Fig. 3B**). Based on UniProt annotation, the experimental construct of T1207-D1 spans two proteins (residues 758-806 and 807-1021) cleaved from this polyprotein. Compared to full biological sequences (residues 758-806 and 807-1021), T1207-D1 is a truncated version mapping to residues 776-919 in the polyprotein (**Fig. 3B** top). We predicted 3D structures of the full sequences using the AF3 web server, and the predicted aligned error (PAE) plot and predicted structure suggest that the biological sequences might fold into two domains (**Fig. 3B** left). The T1207-D1 target contains the entire N-terminal domain and part of the C-terminal helical domain. This truncation probably resulted in an alternative conformation as observed in the experimental structure (**Fig. 3B** middle), where the helical fragment from the C-terminal domain folds back to the N-terminal domain and forms a beta sheet with the N-terminal beta strands. Four out of five models from the Wallner group, as well as a few models from several other groups, resemble the experimental structure, while most other models predicted the segment from the C-terminal domain as an extended helix[29]. We speculate that the experimental structure is an alternative conformation of this protein induced by truncation. Current methods, being insensitive to truncation, might have failed to predict the conformational changes.

Induced fit upon binding: Several most challenging targets from protein complexes contain regions that interact more extensively with other proteins than within the targets. One example is T1267s1-D1 (**Fig. 3C**), a component of the SPARDA complex, which plays an important role in the bacteria defense system. While most groups were able to predict the structured regions of this target, they failed to accurately predict the orientation of the loop mediating its interaction with a binding partner. This loop might undergo induced fit upon binding, and its conformation can only be correctly predicted in the context of its parent oligomeric target H1267, which is a challenging target only correctly predicted by a couple of groups, including the kozakovvajda group. In addition, EUs T1271s1-D1, T1271s5-D1, T1271s8-D1, and T1271s8-D2 are derived from a large RNA editing core complex, and they suffer from similar issues (see the CASP16 website for additional details).

Irregularities in secondary structures: Consistent with previous observations[58,59], we noticed a tendency of over-stabilizing helices or strands in flexible regions, and incorrectly predicting irregular secondary structures such as bent helices, smaller helices (3□□ helix), and larger helical turns (π-helix). Two poorly predicted small EUs, T1228-D3 and T1228-D4, from serine-type integrase SprA contain irregularities in secondary structures and flexible loops connecting short helices. Most predictors missed such irregularities and predicted longer and straight helices instead (**Fig. 3D** and **3E**). This is possibly related to the fact that the training data of AF2/3 are biased towards ordered structures and regular secondary structures. Additionally, these irregularities in secondary structures in a target might not be conserved in its homologs (for example, a homologous target, T1239, does not have these irregularities). Therefore, coevolutionary information derived from MSAs might be misleading as well.

Performance evaluation and ranking

Scoring formula: We investigated a number of classic prediction assessment measures. To understand the relationships among various measures, we conducted correlation analysis and factor analysis (supplemental **Fig. S4**). These analyses allowed us to group the evaluation metrics into categories, from which nine representative measures were selected: **GDT_HA, QSE, reLLG_const, SphGr, CAD_AA, GDC_SC, AL0_P, lDDT,** and **MolProbity score**. In Methods, we briefly describe each score.

Adhering to the CASP protocol, we computed z-scores (see Methods) for each selected measure on models of each EU. To avoid over-penalizing failed or missing predictions, we imputed z-scores below -2 and missing values as -2. An alternative strategy frequently used in CASP is to set z-scores below 0 to 0. We tried both strategies in our assessment and prioritized the former because it gives us better discrimination between groups. In an era when most targets are correctly predicted and the difference in performance between groups is subtle, being able to consistently generate high-quality structure for every target is important.

The total z-score of a prediction is computed as the weighted sum of z-scores of the nine measures. The assigned weights to each measure reflect both precedents from previous CASP experiments and our judgment regarding the importance of each measure. For example, the weight of GDT_HA, QSE, and reLLG_const being 1/6 is consistent with CASP15[30]. The higher weights on these scores emphasize three aspects: the importance of overall structure[55], quality evaluation, and how these models can help experimentalists in molecular replacement[60,61], respectively. The other 6 measures are given either 1/8 or 1/16 weight, depending on their relative importance in estimating model quality.

It is important to note that due to the high correlation between different measures (supplemental **Fig. S4**), the ranking, especially the selection of the winning group, is largely not sensitive to the change of weights. For example, changing the weight of a component score to 0 will result in a very similar ranking, and the top groups will largely remain the same (**Table S1**). Only changing the weight of the Molprobity score to a higher value will significantly change the ranking because Molprobity scores do not correlate with structural similarity between a target and a model, and different groups vary greatly in MolProbity scores of their models. While it is meaningful to use Molprobity scores to encourage people

to mind the geometric accuracy of predicted structures, we chose to use a low weight, as in CASP15, to prevent it from dominating the ranks.

Each monomer target in CASP16 contains one to four EUs, and two targets with the largest number of EUs (T1228 and T1239) share similar 3D structures. To prevent targets with more EUs from dominating the evaluation, we assign a weight to each EU based on the number of EUs per target (i.e., a weight of 1/N for each EU if one target has N EUs), resulting in an equal weight for each target. The final scoring formula is presented below.

$$\begin{aligned}\mathbf{Z(monomer)} = \sum_{\text{target}} \Bigg(\sum_{\text{EU}} w_{\text{EU}} \Big(& \frac{1}{16} \times \big(Z(\text{GDC_SC}) + Z(\text{AL0_P}) + Z(\text{lDDT}) + Z(\text{MolProbity} - \text{Score}) \big) \\ & + \frac{1}{8} \times \big(Z(\text{SphGr}) + Z(\text{CAD}_{\text{AA}}) \big) \\ & + \frac{1}{6} \times \big(Z(\text{GDT_HA}) + Z(\text{QSE}) + Z(\text{reLLG_const}) \big) \Big) \Bigg), w_{\text{EU}} = \mathbf{1/n(EUs)}\end{aligned}$$

Phase 1 ranking: We consider Phase 1 the primary phase of CASP16 due to the following reasons. First, its setup is the most similar to traditional CASP experiments; second, this phase contains the largest number of EUs; third, predictors show the highest participation rate in this phase. We ranked groups by the aforementioned formula over all Phase 1 EUs. Additionally, we conducted a head-to-head comparison with bootstrapping (see Methods) to provide statistical significance for performance differences between groups. The leading groups in Phase 1 are Yang, Yang-Multimer, Yang-Server, MULTICOM, and MULTICOM_human [cite: CASP16 paper from Jianyi Yang and Jianlin Cheng], which significantly outperform the rest (**Fig. 4A** and **4B**). By looking at the contribution of individual scores, we noted that the winners typically excel in measures like GDT_HA and reLLG_const. However, they often score lower in MolProbity, suggesting that their models may violate more physicochemical constraints. We encourage future participants to pay greater attention to the physical validity of their models. Two of the top-performing groups in CASP15, PEZYFoldings and UM-TBM (from Zheng lab), are less competitive in CASP16.

Main ranking based on targets from all phases: CASP16 introduced two new Phases—Phase 0 and Phase 2—which expanded the scope of monomer structure prediction. We combined targets from all three phases to determine the overall ranking for CASP16 monomer prediction and considered this as the main ranking. All three groups from the Yang lab [cite: CASP16 paper from Jianyi Yang] had the highest z-score, which aligns with the results from Phase 1, and their superior performance is statistically significant (**Fig. 4C** and **4D**). This consistency highlights the robustness of their superior performance across prediction Phases in CASP16, and we consider the Yang lab as the winner for monomeric structure prediction.

Phase 0 ranking: Phase 0 is designed to challenge predictors' ability to model oligomeric complexes by not providing stoichiometry information to them. Incorrectly predicted stoichiometry and complex structures might affect the prediction accuracy of monomers derived from these complexes, because interaction with other partners might affect the conformations of these monomers. Phase 0 rankings differ noticeably from those in Phase 1, but surprisingly, groups that performed the best in Phase 0 monomer targets (supplemental **Fig. S5A**) are not the winners of Phase 0 in the oligomer category. Instead, several groups that did not utilize AF3 (NKRNA-s, MIEnsembles-Server, Zheng-Multimer, and Zheng) were ranked at the top for Phase 0 monomer targets. We speculate that this is because the winners (e.g., the Yang lab) of the monomer category relied on extensive use of AF3 to

obtain better models, but they were limited by the quota per user implemented in the AF3 web server. Because some Phase 1 targets' sequences were released during Phase 0, and all Phase 2 targets' sequences were released during Phase 1, they had more time to gather AF3 models for the two phases coming after Phase 0, where they outperformed other teams. They might also have chosen to prioritize Phase 1 targets in modeling via the AF3 web server.

Phase 2 ranking: Phase 2 is designed to evaluate the predictors' ability to utilize the large number of 3D models generated by MassiveFold in their prediction pipeline. However, the integrity of the Phase 2 experiment may have been compromised, as the winning teams likely conducted extensive sampling using their own pipelines in combination with the AF3 web server. The groups from Yang Lab demonstrated an even greater advantage in Phase 2 compared to earlier phases (supplemental **Fig. S5B**), consistent with the observations in the following sections that the groups from Yang Lab are still able to improve in Phase 2. This result may reflect their superior ability to effectively leverage large amounts of AF2/3 models to improve prediction quality or a more extensive usage of AF3 to iteratively refine the input sequences for better model quality [cite: CASP16 paper from Jianyi Yang].

Rankings partitioned by difficulty: In the earliest CASP exercises, there were at least two categories in structure prediction: comparative modeling and *ab initio* (new fold)[1,62,63]. These two categories further became what is known today as template-based modeling and template-free modeling[64,65]. Before the use of coevolution-based contact prediction and the emergence of deep learning methods[66], having homologous structures in the PDB was critical for correctly predicting the fold of a target[67,68]. However, as modern deep learning methods have revolutionized protein structure prediction techniques, having templates detectable by homology search is no longer necessary for obtaining high-quality models. Notably, AF2's performance on template-free targets approached that of easier ones[9], making the traditional dual-category assessment unnecessary. Consequently, since CASP15, the organizers merged these categories into a unified evaluation of single protein modeling[30].

We maintain this unified approach in CASP16. We noticed that there was no significant correlation between the availability of PDB templates (found by HHsuite[69]) and model quality (supplemental **Fig. S6**). However, it remains meaningful to emphasize groups' performance on targets that are challenging. As shown in **Fig. 3A**, targets remaining difficult to the field can be separated from the rest based on GDT_TS95, and we considered targets below the red line in **Fig. 3A** as difficult and the rest as easy. We evaluated groups' performance on the difficult and easy targets separately and the results are shown in supplemental **Fig. S7**. Notably, groups such as Wallner, CSSB-Human, and Yang performed relatively well on difficult targets. In particular, both the Wallner and the CSSB-Human groups were able to correctly predict the topology for the aforementioned challenging target, T1207-D1 (**Fig. 3B**), significantly outperforming other groups on this target. Additionally, CSSB-Human achieved a relatively better performance in MolProbity score, which also contributes to its top ranking in the category of difficult targets.

Insights from additional CASP16 experiments

Comparisons between different phases: In CASP16, each successive phase was designed to provide predictors with additional information with the aim of reducing the

difficulty and stimulating better models. We wondered if CASP16 groups could successfully utilize such additional information, and thus we compared each group's performance difference between consecutive phases (in GDT_HA), focusing on shared targets between them. We focused on 28 groups showing superior performance in our final ranking (targets across 3 phases). For the comparison between Phase 0 and Phase 1, we focused on the first models because the first models reflect each group's best estimate of stoichiometry. It is common for the predictors to use a diverse range of predicted stoichiometry in all five models, and it will be less informative to use the best models here.

Overall, providing the stoichiometry information in Phase 1 and providing over 8,000 MassiveFold models in Phase 2 helped over 60% groups to improve their performance (**Fig. 5A** and **5B**), but such improvement is more noticeable for Phase 1. The smaller improvement in Phase 2 is likely related to two additional factors. First, many groups possibly emphasized more on Phase 1 than the other two phases. Second, the winning groups mostly have their own pipelines to perform extensive sampling through AF2 and AF3, thus, providing additional MassiveFold models is not necessarily helpful; instead, people had more time to sample more AF3 models in Phase 2, which could have been more important. Below, we highlight several cases that are indicative of the differences between phases.

Heavier human involvement in Phase 1 helps to improve model quality: An example of extreme improvement in Phase 1 is between target T1257-D1 and T0257-D1 (**Fig. 5C**). T0257-D1/T1257-D1 is from a trimer of the side tail fiber protein pb1 of bacteriophage. Predicting the stoichiometry of such complexes should be trivial based on AF2/3 prediction and homology-based inference. However, a significant improvement is observed due to the selection of a straight model (**Fig. 5C** left) over a bent one (**Fig. 5C** right) by the MULTICOM group. Both AF3 server and ColabFold have a tendency to bend such elongated proteins and are not likely to generate the correct conformation automatically (supplemental **Fig. S8**). However, knowledge about viral tail proteins indicates that they should typically adopt highly straight and elongated conformations. We speculate that during Phase 0 limited time constrained the predictors to perform heavy manual intervention and model selection. However, since the same target is passed on to Phase 1, predictors may have sufficient time to guide AF2 by providing straight templates and selecting the correct conformation by human intuition.

Knowing stoichiometry helps people to select the correct monomer conformation: Another example of extreme improvement in Phase 1 is between target T0270-D2 and T1270-D2 (**Fig. 5D**). These targets are of HtrA serine protease from *Borreliella burgdorferi*, a pathogen causing Lyme disease. In Phase 0, the KiharaLab group [cite: CASP16 paper from Daisuke Kihara] sampled multiple stoichiometry and selected a 12-mer (**Fig. 5D** bottom left) as their first model. However, this target should be a hexamer instead. In Phase 1, the same group selected a hexamer as model_1 (**Fig. 5D** bottom right). Although this hexameric model is also suboptimal in its oligomeric interfaces, it allowed them to correctly predict the monomer conformation. Comparison between the incorrect Phase 0 prediction (**Fig. 5D** top middle) and the correct Phase 1 prediction (**Fig. 5D** top right) revealed that the former involves domain swapping between interacting chains, resulting in a less compact monomer model.

Extensive sampling improves model quality for a challenging target: The challenging target T1207-D1 (**Fig. 3B**) demonstrates the benefit of extensive sampling. In Phase 1, only very few groups, especially the Wallner group which performs extensive sampling using

AFsample[29,70], were able to correctly predict its structure. However, leveraging the extensive models provided in Phase 2, other groups were able to identify the correct conformation (**Fig. 5E**). This suggests that extensive sampling enabled by MassiveFold can enhance structure prediction accuracy in certain cases.

Evaluation of Model 6: Model 6 predictions are restricted to use MSAs generated by ColabFold, allowing us to separate the contribution by improved MSAs and advanced structure prediction deep learning networks. To assess both the quality of MSAs and the network architectures, we conducted two comparative analyses: 1) model 6 vs. model 1 from the same group – to evaluate the performance of each group's own MSA construction strategy; 2) model 6 vs. ColabFold model – to assess how the group's modeling network performs relative to ColabFold, given the same MSA input. Groups such as Zheng, MULTICOM_human, MULTIFOLD, and PEZYFoldings [cite: CASP16 paper from Zheng, Jianlin Cheng's and PEZYFolding's team] are able to obtain better results for model 1 than model 6, indicating that their group's MSA is better for structure predictions (supplemental **Fig. S9A**). Meanwhile, groups such as Guijunlab-QA, Guijunlab-Human, and MULTICOM [cite: CASP16 paper Guijun Zhang and Jianlin Cheng], were able to outperform ColabFold using the same MSAs, although such an advantage appears to mainly come from several targets that ColabFold struggled with (supplemental **Fig. S9B**).

Although these 6th models were designed to provide invaluable insights about strategies used in the field, they were not considered in the main CASP16 rankings, and consequently, few groups submitted model 6 predictions. Even among those groups that did, many model 6 submissions were missing for a large number of targets. As a result, the available data for model 6 is too limited to support any robust analysis or strong conclusions. We strongly encourage future CASP participants to submit more comprehensive model 6 predictions. By doing so, it would allow for a more systematic and quantitative evaluation of how MSA quality and innovations in model architectures independently contribute to structure prediction performance.

Quality evaluation

As the field advances in modeling monomer structures and large-scale model generation, the ability to rank predicted structures and predict model quality becomes more important in this post-AlphaFold era. We thus performed a series of analyses aiming at evaluating CASP16 groups' ability in self-assessment and model selection.

Ranking by the first model: The first model represents the structure that the predictor is most confident in and thus can serve as an indicator of both prediction quality and model selection ability. Interestingly, if we evaluate the performance of CASP16 groups by the first models, the Yang lab maintains the top position, but the AF3 server rose to second place (**Fig. 6A**), showing a performance not statistically worse than the top group (**Fig. 6B**). Yang's high rank can likely be attributed to its superior modeling accuracy, whereas AF3's increase in rank is likely to be driven by its improvement in model evaluation and ranking capacity.

Frequency of first models being the best models: To directly evaluate each group's ability in selecting the first model, we calculated the percentage of cases in which a group's first model was also the best among the five submitted models for that target. Among the top-ranking groups, many can do only slightly better than a random selection, where the first model has a 20% chance of being the best (**Fig. 6C** and **6D**). One group, PEZYFoldings,

was able to reach a ratio of 30/74 (**Fig. 6D**), highlighting a robust ability in ranking its own models effectively. This is also consistent with their best performance in selecting the first model for oligomeric targets [cite: our oligomer evaluation paper].

Ranking by QSE (a.k.a. ASE) scores: Starting from CASP15, predictors are required to submit their own "pLDDT score" in the b-factor column of their predicted structures[33]. QSE score is subsequently developed to measure the agreement between pLDDT and the real LDDT based on the comparison between a model and a target. By ranking only using QSE scores, we discovered that the AF3 server obtained the highest QSE scores, followed by a couple of other groups, i.e., elofsson and falcon2, that explicitly rely on the AF3 server (**Fig. 6E**). Together with AF3 server's near-top performance in the first model quality, our results suggest the complicated confidence module developed in AF3 does remarkably improve its ability in confidence estimation.

Analysis of MassiveFold results: One key challenge in structure prediction is finding the best models among the diverse structures sampled by modeling with different input MSAs and structure prediction networks. Utilizing the large number of predicted structures by MassiveFold for most CASP16 targets (monomers extracted from protein-nucleotide hybrid complexes did not have MassiveFold results), we were interested in the following question: will we obtain a better prediction among the MassiveFold models if we have a perfect selector to rank them? To address this, we ran LGA for all models from the MassiveFold against their corresponding target. We plotted the best GDT_HA for MassiveFold models on each target against the best models among the around 300 predictions submitted by CASP16 groups for each target.

The best models generated by CASP16 groups are generally better than the best MassiveFold models, suggesting the innovations in modeling strategies used by CASP16 groups resulted in a meaningful improvement in model quality. The most extreme case of human improvement happened to target T1257-D1 discussed above (**Fig. 5C**), a target where automatic predictions from MassiveFold tend to bend the elongated structure. We also noticed that MassiveFold models tend to have worse performance for large targets. Two reasons are likely behind this observation: first, the large complexes are likely harder for automatic prediction without human guidance; second, MassiveFold also generated fewer models (2,040 or 4,040, instead of 8,040) for these large targets due to limits in computational resources.

However, the pool of structures generated by MassiveFold did contain better models for some targets, indicating that our ability to model monomer structures can be further improved by optimizing target selection metrics. Notably, MassiveFold predictions were generated based on AF2-like networks. The diffusion module used by AF3 made it intrinsically suitable for generating a diverse set of predictions and enlarging the pool of possible models to select from. Thus, we expect that further performance improvement can be enabled by better quality assessment and model ranking methods, highlighting self-evaluation as a critical direction for future improvement..

Lack of obvious progress in monomer prediction

To evaluate whether the field has made progress from CASP15 to CASP16, we compared the predictors' performance in CASP16 against CASP15. Because the apparent performance of different CASP experiments will be significantly affected by the overall

difficulty of targets, we need a method that has not changed between the two CASP experiments to serve as an indicator of target difficulty. We used ColabFold's performance as an indicator of target difficulty. ColabFold participated in CASP15 and submitted predictions to all CASP15 targets. Nearly the same ColabFold (colabfold_baseline) pipeline with updated sequence and structure databases was applied to 59 out of 74 CASP16 monomer targets in Phase 1 to help estimate the target difficulties: targets derived from hybrid targets (with nucleotides) cannot be modeled by ColabFold and were excluded. Accordingly, we restricted our comparative analysis to targets with ColabFold predictions.

We computed the average GDT_HA scores of the best models from the top 50 groups in CASP15 and CASP16 (**Fig. 7A**). The highest GDT_HA among CASP participants increased from 71.6 to 77.4. However, the GDT_HA of ColabFold also increased—from 64.7 in CASP15 to 72.2 in CASP16—suggesting a potential shift in the overall difficulty of the targets. The distribution of GDT_HA scores across targets further supports this interpretation, with CASP15 containing a higher proportion of difficult targets compared to CASP16 (**Fig. 7B**). The difference in target difficulty between the two CASP experiments prevents direct comparison of performance, calling for more robust statistical analysis that can take target difficulty levels into account. One obvious difference is a greater number and fraction of groups outperformed ColabFold in CASP16, indicating the field is surpassing AF2, possibly due to the contribution of AF3.

To evaluate the progress in a more robust fashion, we first studied the relationship between target difficulty, as indicated by the GDT_HA score of ColabFold's first model, and the field's optimal performance (in CASP15 or CASP16) on a target (**Fig. 7C**). Differences in target difficulty can confound evaluation of the progress. Easier targets offer limited room for improvement over the ColabFold baseline, whereas harder targets correlate with bigger performance differences between the best model and the ColabFold model (a potential improvement). To more objectively assess progress between CASPs, we implemented two statistical approaches to mitigate biases arising from variation in target difficulty.

First, we applied the Hungarian algorithm[71], which finds the optimal one-to-one matching between two sets, to select a subset of CASP15 targets to match the difficult level of CASP16 targets as much as possible. In our case, the set of matched targets (connected by the red lines in **Fig. 7D**) shows very similar overall difficulty levels between CASP15 and CASP16, with average ColabFold GDT_HA of 72.06 and 72.23, respectively. Based on this matching to control the shift in target difficulty, we calculated the performance improvement over ColabFold (as the difference between best GDT_HA and ColabFold's GDT_HA) in CASP15 and CASP16 (**Fig. 7E**) and discovered that the performance in CASP15 and CASP16 is very similar (mean improvement over ColabFold being 8.73 and 8.78 by GDT_HA) without any statistical significance in performance difference (p-value = 0.92).

Second, we used a weighted bootstrap technique to generate samples (with replacement) of CASP15 targets while trying to make each bootstrap sample have a similar ColabFold GDT_HA distribution as CASP16 targets. For each bootstrap sample of CASP15 targets, we computed their average ColabFold GDT_HA (target difficulty) and the average difference in best model's GDT_HA compared to ColabFold's (performance improvement over ColabFold baseline) in CASP15, and the distribution of these two measurements are shown in **Fig. 7F**. Among these bootstrap samples, we observed a clear negative correlation between overall target difficulty and performance improvement. Among CASP15 bootstrap samples with similar overall difficulty levels (near the red dot representing CASP16), about 1/3 show better

overall performance than CASP16. The average performance improvement among these CASP15 bootstrap samples is 8.35 GDT_HA units, slightly lower than the observed performance improvement in CASP16 (8.78 GDT_HA units). However, given that these numbers are so close to each other, we do not think they support meaningful progress from CASP15 to CASP16.

Given AF3 allows people to include small molecules in their models of proteins, we wondered whether using AF3 allowed CASP16 groups to predict atomic details around the functional sites (such as metal and small molecule binding sites) to higher accuracy. We thus focused on residues around metals and biological ligands found in structures submitted to the prediction center by experimental labs. We evaluated the prediction accuracy of side chains of these ligand-binding residues using GDC_all[55] and compared the performance between CASP15 and CASP16 (supplemental **Fig. S10** and **S11**). Again, we did not notice any remarkable improvement in the modeling accuracy for these important functional sites.

Inter-domain interaction evaluation

Inter-EU interaction prediction was conducted by monomer or assembly assessors in the previous CASPs[7,72]. Before the recent breakthrough in protein structure modeling, the community struggled in predicting domain-domain interactions (DDIs) without homologous PDB templates. Thus, a EU in early CASPs frequently corresponds to a domain. As the community's ability to model protein structures continues to grow[7,72–75], the accuracy in modeling DDIs has drastically increased. In recent years, multiple interacting domains have been frequently merged into a single EU, representing a unit whose structure can be predicted as well as single domains by the community[39]. With the improvement in modeling inter-domain interactions, the size of EUs in CASP also increased.

Among CASP16 monomer targets, we identified 63-68 pairs of DDIs based on Chainsaw (for domain parsing) and PISA (for interface analysis). The uncertainty in the number of DDIs is due to some proteins (T1228 and T1239) adopting multiple conformations and showing variability in inter-domain interactions. 50 (74%–79 %) of these interacting domain pairs were included in the same EU, suggesting that the majority of DDIs can be accurately modeled by the field. The remaining 13-18 DDIs were partitioned into different EUs, and it is not clear whether their relative orientations are intrinsically flexible or they represent challenging cases for DDI modeling.

To probe whether these inter-EU DDIs are of biological relevance, we used PISA[76], a well-established software that judges whether an observed interface in an experimental structure is biologically relevant, to assess the interfaces between these inter-EU DDIs. As a control, we also used PISA to evaluate the interfaces between DDIs within the same EU. Plotting the interface sizes versus PISA P-value (a measure of biological relevance of an interface), we found that inter-EU DDIs tend to have smaller interface sizes and higher PISA P-value. A PISA P-value of greater than 0.5 is suggested to be indicative of experimental artifacts instead of biologically relevant interactions. Only 5 inter-EU DDIs show a PISA P-value < 0.5, suggesting that they could be biologically relevant and suitable targets for evaluating DDIs. However, manual inspection suggests only three of them are not suitable targets for DDI prediction due to different reasons.

In particular, two EUs, T1239-D2 and T1239-D4 (orange circle in **Fig. 8A** and **Fig. 8B**), might interact or not according to different monomers within the same protein complex,

suggesting the flexibility of their relative orientation. Also, the interaction between T1295-D1 and T1295-D2 (red circle in **Fig. 8A** and **Fig. 8C**) was known from an experimental structure (PDB id: 6E14) showing 100% sequence identity to the target, which was solved before CASP16. Finally, two "interacting" EUs, T1269-D2 and T1269-D3 (magenta circle in **Fig. 8A** and **Fig. 8D**), are connected by a small domain. The boundary between T1269-D2 and T1269-D3 was incorrectly placed in the middle of the small domain due to errors in domain parsing. The observed "interaction" between them only happens within this small domain, and outside this small domain, the two EUs do not interact.

Only the two interfaces formed by T1218-D1 and T1218-D2, and T1218-D2 and T1218-D3 (cyan circle in **Fig. 8A** and **Fig. 8E**) are likely to be biologically meaningful and suitable for DDI prediction and evaluation. Therefore, we reason that the evaluation of inter-EU interaction became unnecessary in this CASP, given that the significant interactions are already within each EU, and they are all correctly predicted by the top predictions (supplemental **Fig. S1**)

Conclusions and Future Perspectives

Single-domain protein structure prediction is now largely a solved problem, as evidenced by CASP16, where not a single fold was incorrectly predicted. However, specific challenges persist that continue to test the limits of current methods. Larger proteins, as well as viral and eukaryotic proteins, and those with shallow multiple sequence alignments (MSAs), remain more difficult to model accurately. Structural features such as truncated regions, irregular secondary structures, and interchain regions are also prone to prediction errors. These cases highlight important directions for further development within the prediction community.

As in CASP15, the top-performing strategies in CASP16 relied heavily on AlphaFold2 (AF2) and AlphaFold3 (AF3). Our analysis of Phase 2 predictions and the MassiveFold results confirms that sampling more structures has the potential to improve outcomes. Additionally, constructing better MSAs is likely to enhance prediction accuracy. However, due to the limited submissions for the "Model 6" trial, the benefit of improved MSA quality has not yet been rigorously benchmarked. We encourage broader participation in Model 6 in future CASP experiments, as it provides a valuable opportunity to assess MSA quality independently of network architecture.

Overall, performance on monomer targets in CASP16 showed minimal improvement compared to CASP15. This is consistent with the observation that simply increasing AF2 sampling often yields models comparable to or better than those produced by AF3. That said, AF3 demonstrated superior accuracy in confidence estimation, and further optimization following its full release could lead to new breakthroughs in the field.

Community engagement in the additional CASP16 challenges—Phase 0, Phase 2, and Model 6—was notably lower than in the traditional Phase 1 challenge. Nevertheless, we believe these newer challenges are critical for driving future progress. Predicting complex structures without prior stoichiometry information, as in Phase 0, represents an important frontier and should be considered as a core component of future CASP evaluations. Phase 2 aimed to stimulate development in model ranking, but its current design conflates model

generation and ranking. A more meaningful approach would be to ask predictors to select their top five models from a fixed set provided by organizers. To encourage participation in Model 6, organizers and future assessors could consider including it in the main performance evaluation, using each group's best model as the basis for comparison. This approach would incentivize submissions of model 6 by giving participating groups additional opportunities to achieve higher rankings.

The field of protein structure prediction has witnessed remarkable progress in recent years, largely driven by the introduction of deep learning methods. Looking ahead, we hope this manuscript can provide a comprehensive evaluation of the current status of protein structure prediction while also highlighting the remaining challenges that could be focused on to continue advancing the field.

Methods

Definition of evaluation units (EUs)

The original EUs were defined according to the CASP15 principles[39] and further manually examined. We made some adjustments based on the following criteria: 1. Merging EUs. We studied whether any pairs of EUs from a target should be merged based on the Grishin plot—initially proposed by Grishin and colleagues in CASP9[77]. If merging two EUs does not significantly affect the performance of top top-ranking groups, we would merge them. 2. Consolidating targets with different conformations. In this CASP, several EUs (T1214-D1, T1228-Ds, T1239-Ds, T1249-D1, T1294-D1) have multiple versions due to different conformations and ligand-induced fit. We evaluated the structural variations between different versions, and when such differences, as measured by RMSD, were above 1□, we used different versions in our score calculation. Otherwise, we used the version 1 designated by the Prediction Center [cite: CASP infrastructure paper by Kryshtafovych et al.]. Different versions of a target were evaluated collectively: model comparisons were conducted against all available conformations, and the highest score obtained from any version was used for the assessment. However, for simplicity and consistency in structural visualizations, only v1 was used in all figures and presentations. 74 EUs were thus defined for Phase 1, and we propagated these definitions to the corresponding targets in Phase 0 and Phase 2, yielding a total of 145 EUs combining all 3 Phases.

Heatmap Generation and Clustering

The highest GDT_HA from each group for each model was collected. Groups submitting less than 25% of the targets were removed. Ward clustering algorithm was applied to predictors and targets based on performance similarity. Clustering algorithms typically require a definition of distance, and in this case, missing values were assigned 50 for simplicity. The orders in heatmaps reflect the results of clustering and do not imply ranking. The annotations describing predictor strategies and target characteristics were derived from participant abstracts and official CASP classifications, respectively.

Performance evaluation and ranking

As mentioned in the main text, we included nine scores, i.e., GDT_HA, AL0_P, CAD_AA, GDC_SC, SphGr, lDDT, reLLG_const, QSE, and MolProbity score. These scores evaluate model accuracy across multiple dimensions: global fold correctness, local geometric precision, and suitability as templates for experimental studies and other applications. Here we briefly describe each score. **GDT_HA**[55]: Measures the percentage of correctly aligned residues after structural alignment, emphasizing high-accuracy similarity in structures; **QSE**: Measures the difference between pLDDT[6] and lDDT[78], serving as a measure of self-assessment; **reLLG_const**[60]: Evaluates the utility of models in molecular replacement (MR) by estimating the log-likelihood gain, a value in MR that measures how much better the experimental data can be predicted with a model than with a random distribution of the same atoms. This score highlights the usefulness of the model for structural biologists; **SphGr**[79]: Evaluates protein structure models by comparing local spherical regions of the model with the target at multiple scales; **CAD_AA**[80]: Evaluates protein model accuracy at an atomic level by comparing the computed residue contact areas between the model and target; **GDC_SC**[55]: Measures the percentage of correctly aligned sidechain atoms; **AL0_P**[55]: Represents the percentage of residues correctly aligned in the model based on the LGA superimposition; **lDDT**[78,81]: Evaluates model-target local distance differences without structural superposition; **MolProbity score**[54]: a comprehensive measure that evaluates how well the model follows physicochemical rules.

Following the tradition of CASPs, we used Z-scores for performance evaluation and ranking. Prior to Z-score calculations, we identified cases where certain models could not be scored on specific metrics due to intrinsic model limitations. In these instances, missing values were imputed using the lowest observed value across all models for the target in that metric. Z-score calculations were carried out in two rounds. In the first round, models with an initial Z-score below −2 were considered outliers and excluded. Z-scores were then recalculated, and for ranking purposes, any Z-scores below −2 were imputed as −2. This approach ensures that groups employing more experimental or speculative methods, which may occasionally result in poor models, are not disproportionately penalized and can still be recognized for consistently strong performance on other targets.

As in previous CASP assessments, final rankings were based on the sum of Z-scores across all targets, emphasizing overall performance. To address concerns that some models may capture the global fold well while failing in other aspects, the composite scoring scheme was used. We made rankings on both first and best models (the model with the highest Z scores of each group). If the first model was not designated by the group, the model with the lowest model ID was used as the first model.

To assess the robustness of the rankings and evaluate whether performance differences between groups are statistically significant, we performed head-to-head comparisons for the top 25% groups based on the cumulative Z-score-based ranking. In such head-to-head comparisons, we compare the performance of each two groups using only the targets for which both groups submitted predictions. We generated 1,000 bootstrap replicates of these common targets to estimate statistical significance. For each replicate, we compared the

groups' performances based on the summed Z-score over the nine component scores and counted how many times each group outperformed the other across the replicates.

Correlation analysis and Exploratory Factor Analysis (EFA)

To explore underlying relationships between these metrics, we conducted correlation analysis and EFA. The correlation matrix is computed with raw scores across evaluation measures. For EFA, we used the factor_analyzer package in Python[82]. The rotation parameter is 'promax', and the method parameter is 'minres'. Note that EFA does not guarantee a unique solution, so the results for EFA are for exploratory purposes rather than a rigorous derivation of formulas.

Model 6 evaluation and cross-phase comparison

Model 6 against model 1 evaluation: For each group, in every target where both model 1 and model 6 were available, we computed the ratio of their GDT_HA scores. Each group's score was calculated as the average ratio across all such targets.

Model 6 against colabfold_baseline model 1 evaluation: For each group, in every target where both this group's model 6 and colabfold_baseline model 1 were available, we computed the ratio of their GDT_HA scores. Each group's score was calculated as the average ratio across all such targets.

Cross-phase comparison: To compare performance across phases, we focused on targets common to both phases under comparison: Phase 0 vs. Phase 1 and Phase 1 vs. Phase 2. For each group, we calculated the gain or loss in GDT_HA by computing the difference between the next phase and a previous one on a per-target basis and then averaging the values across all shared targets.

Quality evaluation

Using the above Performance evaluation and ranking pipeline, QSE ranking was conducted using only the QSE score, and only the first model. To evaluate the potential improvement we can achieve by selecting the best model among the large amount of MassiveFold models, we computed GDT_HA for MassiveFold models using LGA with the following parameters: lga -3 -ie -o1 -sda -d:4 -dgc_sc. The highest score among all MassiveFold GDT_HA was taken. For each target, we compared the best MassiveFold GDT_HA against the scores for the best model, the first model, the first model of the best group, AF3 server model, and colabfold_baseline model. We also recomputed the GDT_HA scores for these CASP models using our version of LGA, and the results were identical to the values provided by the Prediction Center.

Progress evaluation

The same scoring procedure was applied for CASP15 and CASP16 results. The highest GDT_HA from each group for each target was collected. Groups submitting less than 25% of the targets were removed. Targets that are missing from colabfold_baseline were excluded.

Missing values were imputed with the corresponding score from colabfold_baseline, given the assumption that ColabFold is the tool that every predictor can at least use. Average scores were computed using the above data.

To evaluate performance near functional and biologically relevant regions, we defined them as the residues that are within 5 Å distance to any metals or biologically relevant molecules. The residues were taken from the PDB files submitted by experimentalists and mapped to the processed target PDB. The scores were taken from per-residue GDC_all scores from the output of LGA software provided by the organizers. An average score was computed for each target. The scores were then processed in the same way that GDT_HA data is processed. colabfold_baseline was used as a benchmark.

To compare performance between CASP15 and CASP16, we used two statistical methods to select a set of CASP15 targets showing a similar difficulty level distribution to CASP16 targets. We estimated the difficulty of a target by the GDT_HA score of colabFold_baseline's first model, and we used a Hungarian algorithm to generate a one-to-one mapping between CASP15 and CASP16 targets, while minimizing the difficulty difference between members in the two sets. To introduce randomness into this process, we first used Gaussian Kernel Density Estimation to fit a continuous probability distribution for CASP15 and CASP16 target difficulties (by colabfold_baseline's GDT_HA scores). Then, a weighted resampling based on the ratio between the two distributions is done to ensure the bootstrap sample from CASP15 resembles CASP16 in target difficulties.

Inter-EU and Intra-EU evaluation

For EUs that come from the same target, the interaction between each pair of EUs was computed using PISA to obtain inter-EU interaction results. Within each EU, Chainsaw[37] was first applied to parse them into domains. For each EU having more than one domain, the interaction between each pair of domains was computed using PISA to obtain intra-EU interaction results. PISA was used with default parameters. 5% of the residues were truncated on both ends of the EU or domain sequences before running PISA to minimize the influence by the fact that some EU or domain pairs are neighboring in sequence and are thus covalently connected.

Author Contributions

RY and JZhang conducted the assessment with supervision from QC and NVG, and with instructions from AK. RDS and JZhou participated in data analysis and provided suggestions for analysis methods. RY and QC prepared the figures and drafted the manuscript. RY, JZhang, QC, AK, and NVG edited the manuscript.

Acknowledgements

QC is supported by the Endowed Scholars Program in UTSW and I-2095-20220331 from Welch Foundation. NVG is supported by I-1505 from Welch Foundation and 2224128 from National Science Foundation Division of Biological Infrastructure. AK is supported by NIGMS GM100482. RDS is supported by NIGMS GM147367, JZhou is supported by RR190071 from the Cancer Prevention and Research Institute of Texas, and JZhang is supported by NIAID K99AI180984-01A1. The authors thank: All CASP organizers, in particular John Moult, for giving us valuable guidance during CASP evaluation; Gabriel Studer for fixing technical issues in OST software, which help us to gain more insights in the definition of monomer evaluation units; Guillaume Brysbaert for providing MassiveFold structures; Arne Elofsson for providing feedback and pointing out several missing models during our evaluation process.

Declaration of interests

The authors have no competing interests to declare.

Data availability and code availability

CASP16's official website is https://www.predictioncenter.org/casp16/index.cgi. Raw results can be found at https://predictioncenter.org/casp16/results.cgi?tr_type=regular. Processed data, materials, and code are available upon request.

Figure legends

Figure 1. A heatmap of the targets and participating groups. The value in each box is the highest GDT_HA value of one group for one target (Grey indicates that such score is not available). The annotations on the x-axis and y-axis represent the different features of groups and targets, respectively. Only groups that submitted more than 25% of the models are included.

Figure 2. Comparisons of different strategies and correlations between prediction quality and target features. **A)** Of all the groups that provided abstracts and submitted monomer targets, the percentage of AF2 and AF3 usage. **B-D)** y-axis: mean GDT_HA of each target from a subset of groups. x-axis: the partition of groups based on the usage of AF2/3, customized MSA, and PLM. **E-I)** y-axis: 95% quantile of GDT_HA scores for models of a target. X-axis: the partition of targets based on sequence diversity in MSAs, taxonomy, experimental technique, MolProbity quality, and the type of experimental structure it is from.

Figure 3. A) 95% quantile of GDT_TS scores against target size. The red line is our manual

partition of difficulty level. Targets below the line are considered as difficult targets. **B)** Case study for T1207-D1. Top, mapping of the target sequence to its source viral polyprotein. Very few models are correct for this target, possibly due to conformation changes upon sequence truncation. **C)** Case study for T1267s1-D1: predicting its structure requires modeling the whole complex correctly because part of it involves interaction with other subunits. **D)** Case study for T1228-D3: an example of over-stabilizing secondary structures. **E)** Case study for T1228-D4: an example of irregular secondary structures.

Figure 4. A) Ranking using all Phase 1 targets and best models. **B)** Results of head-to-head comparisons with 1,000 bootstrapping samples. The values in each box reflect the percentage of bootstrap samples where a group (labeled on the Y-axis) outperforms another one (labeled on the X-axis). Red font in each black cell indicates significance at a confidence level of 0.05. Bootstrapping is done only between the groups present in **A** and is ranked based on the row sum. **C)** Ranking using all Phase 0, 1, 2 targets and best models. **D)** Similar to B, but done on targets from all phases and using groups present in **C**.

Figure 5. A) Average change of GDT_HA for all the targets passed from Phase 0 to Phase 1. The number on each bar indicates the number of targets present in both Phases submitted by this group. **B)** Average change of GDT_HA for all the targets passed from Phase 1 to Phase 2. The number on each bar indicates the number of targets present in both Phases submitted by this group. **C)** Left: a correct model in Phase 1; right: an incorrect model that bends this long protein. **D)** Top: target and model structure for T0270-D2 and T1270-D2; bottom: the predicted models in their oligomeric context. **E)** Target and model structure for T0270-D2 and T1270-D2.

Figure 6. A) Ranking using Phase 1 targets and first models. **B)** Results of head-to-head comparisons with 1,000 bootstrapping samples. The values in each box reflect the percentage of bootstrap samples where a group (labeled on the Y-axis) outperforms another one (labeled on the X-axis). Red font in each black cell indicates significance at a confidence level of 0.05. Bootstrapping is done only between the groups present in **A** and is ranked based on the row sum. **C)** Statistics on the frequency of the first model having the highest GDT_HA across all models for every target. Only the top 28 groups in the main ranking are presented. **D)** Ranking based on the count of green cells (targets where the first model is the best one for a group) in each row of **C**. **E)** Ranking using only QSE scores. **F)** Values of the highest GDT_HA among MassiveFold models or different sets of models submitted by CASP16 groups. Targets are ordered by the highest GDT_HA from left to right.

Figure 7. A) Average GDT_HA for different groups over all targets submitted by colabfold_baseline for CASP15 (top) and CASP16 (bottom). **B)** Distribution of CASP15 (top) and CASP16 (bottom) colabfold_baseline's best GDT_HA. **C)** Best GDT_HA among CASP15 and CASP16 groups and colabfold_baseline's best GDT_HA. **D)** Mapping CASP15 models to CASP16 models to minimize the difference in colabfold_baseline's GDT_HAs (target difficult measure) between the two CASP experiments. **E)** Average difference of best GDT_HA and colabfold_baseline GDT_HA for the targets in the 1-on-1 mapping in **D**. **F)**

Performance comparisons between CASP16 targets and weighted bootstrap samples of CASP15 targets. These bootstrap samples of CASP15 targets show similar difficulty levels as CASP16 targets (by colabfold_baseline GDT_HA, x-axis) and improvement over colabfold_baseline by best CASP15 or CASP16 models (y-axis).

Figure 8. A) PISA interface size versus p-value for biological relevance for intra- and inter-EU DDIs. Each dot represents a DDI and inter-EU DDIs showing PISA p-value below 0.5 are marked by circles of different colors, and these colors link these dots to the corresponding structures shown in **B-E. B)** A target contains multiple copies of the same chain adopting different conformations. In one possible conformation, T1239_D2 (magenta on the right) and T1239_D4 (cyan on the top) do not interact; in another conformation, T1239_D2 and T1239_D4 interact. **C)** Two EUs from this target, T1295_D1 and T1295_D2, are interacting, but the 3D structures of these two EUs are available in PDB prior to CASP16. **D)** Two EUs appear to "interact" with each other due to a domain boundary issue. T1269_D2 and T1269_D3 are connected to each other in sequence, and the boundary between them is within a smaller domain (in the black circle). The two EUs do not interact with each other except in this small domain. **E)** Three EUs from T1218 with two biologically meaningful interfaces: between T1218_D1 and T1218_D2, and between T1218_D2 and T1218_D3.

References

1. Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins Struct. Funct. Bioinforma.* **23**, (1995).

2. Sussman, J. L. *et al.* Protein Data Bank (PDB): Database of Three-Dimensional Structural Information of Biological Macromolecules. *Acta Crystallogr. D Biol. Crystallogr.* **54**, 1078–1084 (1998).

3. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

4. Moult, J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.* **15**, 285–289 (2005).

5. Jumper, J. *et al.* Applying and improving AlphaFold at CASP14. *Proteins Struct. Funct. Bioinforma.* **89**, 1711–1721 (2021).

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

7. Schaeffer, R. D., Kinch, L., Kryshtafovych, A. & Grishin, N. V. Assessment of domain interactions in the fourteenth round of the Critical Assessment of Structure Prediction (CASP14). *Proteins Struct. Funct. Bioinforma.* **89**, 1700–1710 (2021).

8. Pereira, J. *et al.* High-accuracy protein structure prediction in CASP14. *Proteins Struct. Funct. Bioinforma.* **89**, 1687–1699 (2021).

9. Kinch, L. N., Pei, J., Kryshtafovych, A., Schaeffer, R. D. & Grishin, N. V. Topology evaluation of models for difficult targets in the 14th round of the critical assessment of protein structure prediction (CASP14). *Proteins Struct. Funct. Bioinforma.* **89**, 1673–1686 (2021).

10. Akdel, M. *et al.* A structural biology community assessment of AlphaFold2 applications. *Nat. Struct. Mol. Biol.* **29**, 1056–1067 (2022).

11. Yang, Z., Zeng, X., Zhao, Y. & Chen, R. AlphaFold2 and its applications in the fields of biology and medicine. *Signal Transduct. Target. Ther.* **8**, 1–14 (2023).

12. Durham, J., Zhang, J., Humphreys, I. R., Pei, J. & Cong, Q. Recent advances in predicting and modeling protein–protein interactions. *Trends Biochem. Sci.* **48**, 527–538 (2023).

13. Humphreys, I. R. *et al.* Computed structures of core eukaryotic protein complexes. *Science* **374**, eabm4805 (2021).

14. Fontana, P. *et al.* Structure of cytoplasmic ring of nuclear pore complex by integrative cryo-EM and AlphaFold. *Science* **376**, eabm9326 (2022).

15. Burke, D. F. *et al.* Towards a structurally resolved human protein interaction network. *Nat. Struct. Mol. Biol.* **30**, 216–225 (2023).

16. Durairaj, J. *et al.* Uncovering new families and folds in the natural protein universe. *Nature* **622**, 646–653 (2023).

17. Barrio-Hernandez, I. *et al.* Clustering predicted structures at the scale of the known protein universe. *Nature* **622**, 637–645 (2023).

18. Bennett, N. R. *et al.* Improving de novo protein binder design with deep learning. *Nat. Commun.* **14**, 2625 (2023).

19. Kreitz, J. *et al.* Programmable protein delivery with a bacterial contractile injection system. *Nature* **616**, 357–364 (2023).

20. Yu, J. *et al.* A replisome-associated histone H3-H4 chaperone required for epigenetic inheritance. *Cell* **187**, 5010-5028.e24 (2024).

21. Deneke, V. E. *et al.* A conserved fertilization complex bridges sperm and egg in vertebrates. *Cell* **187**, 7066-7078.e22 (2024).

22. Lyu, J. *et al.* AlphaFold2 structures guide prospective ligand discovery. *Science* **384**, eadn6354 (2024).

23. Zheng, W., Wuyun, Q., Freddolino, L. & Zhang, Y. Integrating deep learning, threading alignments, and a multi-MSA strategy for high-quality protein monomer and complex structure prediction in CASP15. *Proteins Struct. Funct. Bioinforma.* **91**, 1684–1703 (2023).

24. Peng, Z., Wang, W., Wei, H., Li, X. & Yang, J. Improved protein structure prediction with trRosettaX2, AlphaFold2, and optimized MSAs in CASP15. *Proteins Struct. Funct. Bioinforma.* **91**, 1704–1711 (2023).

25. Oda, T. Improving protein structure prediction with extended sequence similarity searches and deep-learning-based refinement in CASP15. *Proteins Struct. Funct. Bioinforma.* **91**, 1712–1723 (2023).

26. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

27. Raouraoua, N. *et al.* MassiveFold: unveiling AlphaFold's hidden potential with optimized and parallelized massive sampling. *Nat. Comput. Sci.* **4**, 824–828 (2024).

28. Wallner, B. Improved multimer prediction using massive sampling with AlphaFold in CASP15. *Proteins Struct. Funct. Bioinforma.* **91**, 1734–1746 (2023).

29. Wallner, B. AFsample: improving multimer prediction with AlphaFold using massive sampling. *Bioinformatics* **39**, btad573 (2023).

30. Simpkin, A. J. *et al.* Tertiary structure assessment at CASP15. *Proteins Struct. Funct. Bioinforma.* **91**, 1616–1635 (2023).

31. Zheng, W. *et al.* Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data. *Nat. Methods* **21**, 279–289 (2024).

32. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).

33. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XV. *Proteins Struct. Funct. Bioinforma.* **91**, 1539–1549 (2023).

34. Xu, Y., Xu, D. & Gabow, H. N. Protein domain decomposition using a graph-theoretic approach. *Bioinformatics* **16**, 1091–1104 (2000).

35. Zhou, H., Xue, B. & Zhou, Y. DDOMAIN: Dividing structures into domains using a normalized domain–domain interaction profile. *Protein Sci.* **16**, 947–955 (2007).

36. Postic, G., Ghouzam, Y., Chebrek, R. & Gelly, J.-C. An ambiguity principle for assigning protein structural domains. *Sci. Adv.* **3**, e1600552 (2017).

37. Wells, J. *et al.* Chainsaw: protein domain segmentation with fully convolutional neural networks. *Bioinformatics* **40**, btae296 (2024).

38. Kinch, L. N., Schaeffer, R. D., Kryshtafovych, A. & Grishin, N. V. Target classification in the 14th round of the critical assessment of protein structure prediction (CASP14). *Proteins Struct. Funct. Bioinforma.* **89**, 1618–1632 (2021).

39. Kryshtafovych, A. & Rigden, D. J. To split or not to split: CASP15 targets and their processing into tertiary structure evaluation units. *Proteins Struct. Funct. Bioinforma.* **91**, 1558–1570 (2023).

40. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

41. Haghani, M., Bhattacharya, D. & Murali, T. M. NEFFy: A Versatile Tool for Computing the Number of Effective Sequences. 2024.12.01.625733 Preprint at https://doi.org/10.1101/2024.12.01.625733 (2024).

42. Chang, L. & Perez, A. AlphaFold2 knows some protein folding principles. *bioRxiv* 2024.08.25.609581 (2024) doi:10.1101/2024.08.25.609581.

43. Kemena, C. & Notredame, C. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics* **25**, 2455–2465 (2009).

44. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).

45. Koonin, E. V., Makarova, K. S. & Aravind, L. Horizontal Gene Transfer in Prokaryotes: Quantification and Classification1. *Annu. Rev. Microbiol.* **55**, 709–742 (2001).

46. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).

47. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).

48. Hou, Q. *et al.* Using metagenomic data to boost protein structure prediction and discovery. *Comput. Struct. Biotechnol. J.* **20**, 434–442 (2022).

49. Koonin, E. V., Dolja, V. V. & Krupovic, M. The logic of virus evolution. *Cell Host Microbe* **30**, 917–929 (2022).

50. Jayaraman, B., Smith, A. M., Fernandes, J. D. & Frankel, A. D. Oligomeric viral proteins: small in size, large in presence. *Crit. Rev. Biochem. Mol. Biol.* **51**, 379 (2016).

51. Kim, R. S., Levy Karin, E., Mirdita, M., Chikhi, R. & Steinegger, M. BFVD—a large repository of predicted viral protein structures. *Nucleic Acids Res.* **53**, D340–D347 (2025).

52. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).

53. Wüthrich, K. NMR with Proteins and Nucleic Acids. *Europhys. News* **17**, 11–13 (1986).

54. Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35**, W375–W383 (2007).

55. Zemla, A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31**, 3370–3374 (2003).

56. Buel, G. R. & Walters, K. J. Can AlphaFold2 predict the impact of missense mutations on structure? *Nat. Struct. Mol. Biol.* **29**, 1–2 (2022).

57. Woods, H. *et al.* Computational modeling and prediction of deletion mutants. *Structure* **31**, 713-723.e3 (2023).

58. Laurents, D. V. AlphaFold 2 and NMR Spectroscopy: Partners to Understand Protein Structure, Dynamics and Function. *Front. Mol. Biosci.* **9**, (2022).

59. Wilson, C. J., Choy, W.-Y. & Karttunen, M. AlphaFold2: A Role for Disordered Protein/Region Prediction? *Int. J. Mol. Sci.* **23**, 4591 (2022).

60. Millán, C. *et al.* Assessing the utility of CASP14 models for molecular replacement. *Proteins Struct. Funct. Bioinforma.* **89**, 1752–1769 (2021).

61. Read, R. J. & Chavali, G. Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins Struct. Funct. Bioinforma.* **69**, 27–37 (2007).

62. Mosimann, S., Meleshko, R. & James, M. N. G. A critical assessment of comparative molecular modeling of tertiary structures of proteins. *Proteins Struct. Funct. Bioinforma.* **23**, 301–317 (1995).

63. Defay, T. & Cohen, F. E. Evaluation of current techniques for Ab initio protein structure prediction. *Proteins Struct. Funct. Bioinforma.* **23**, 431–445 (1995).

64. Kopp, J., Bordoli, L., Battey, J. N. D., Kiefer, F. & Schwede, T. Assessment of CASP7 predictions for template-based modeling targets. *Proteins Struct. Funct. Bioinforma.* **69**, 38–56 (2007).

65. Jauch, R., Yeo, H. C., Kolatkar, P. R. & Clarke, N. D. Assessment of CASP7 structure predictions for template free targets. *Proteins Struct. Funct. Bioinforma.* **69**, 57–67 (2007).

66. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP) — round x. *Proteins Struct. Funct. Bioinforma.* **82**, 1–6 (2014).

67. Huang, Y. J., Mao, B., Aramini, J. M. & Montelione, G. T. Assessment of template-based protein structure predictions in CASP10. *Proteins Struct. Funct. Bioinforma.* **82**, 43–56 (2014).

68. Tai, C.-H., Bai, H., Taylor, T. J. & Lee, B. Assessment of template-free modeling in CASP10 and ROLL. *Proteins Struct. Funct. Bioinforma.* **82**, 57–83 (2014).

69. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).

70. Kalakoti, Y. & Wallner, B. AFsample2: Predicting multiple conformations and ensembles with AlphaFold2. 2024.05.28.596195 Preprint at https://doi.org/10.1101/2024.05.28.596195 (2024).

71. Kuhn, H. W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **2**, 83–97 (1955).

72. Guzenko, D., Lafita, A., Monastyrskyy, B., Kryshtafovych, A. & Duarte, J. M. Assessment of protein assembly prediction in CASP13. *Proteins Struct. Funct. Bioinforma.* **87**, 1190–1199 (2019).

73. Kosciolek, T. & Jones, D. T. Accurate contact predictions using covariation techniques and machine learning. *Proteins Struct. Funct. Bioinforma.* **84**, 145–151 (2016).

74. Wang, S., Sun, S. & Xu, J. Analysis of deep learning methods for blind protein contact prediction in CASP12. *Proteins Struct. Funct. Bioinforma.* **86**, 67–77 (2018).

75. Kandathil, S. M., Greener, J. G. & Jones, D. T. Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. *Proteins Struct. Funct. Bioinforma.* **87**, 1092–1099 (2019).

76. Krissinel, E. & Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **372**, 774–797 (2007).

77. Kinch, L. N. *et al.* CASP9 target classification. *Proteins Struct. Funct. Bioinforma.* **79**, 21–36 (2011).

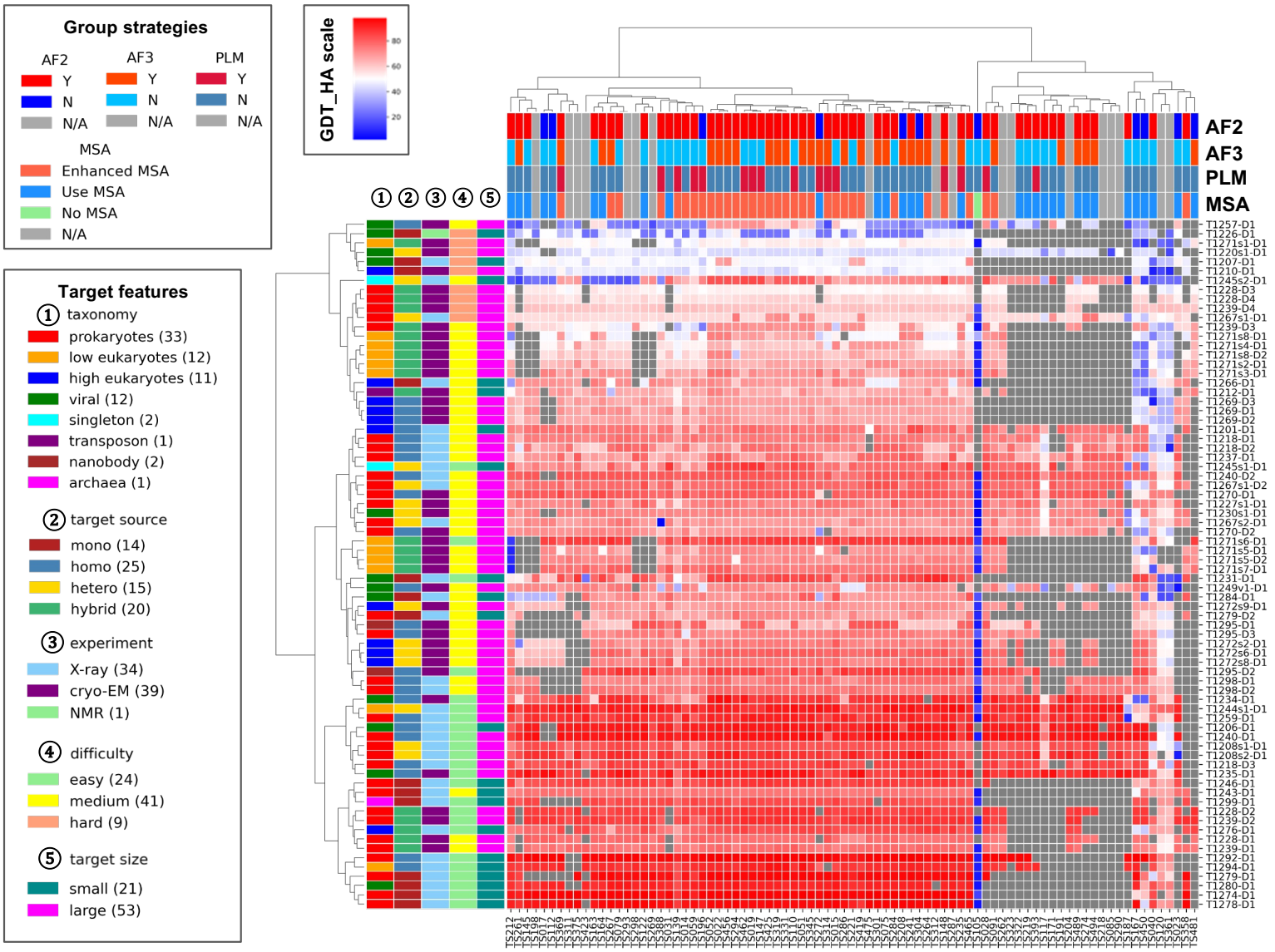
78. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).

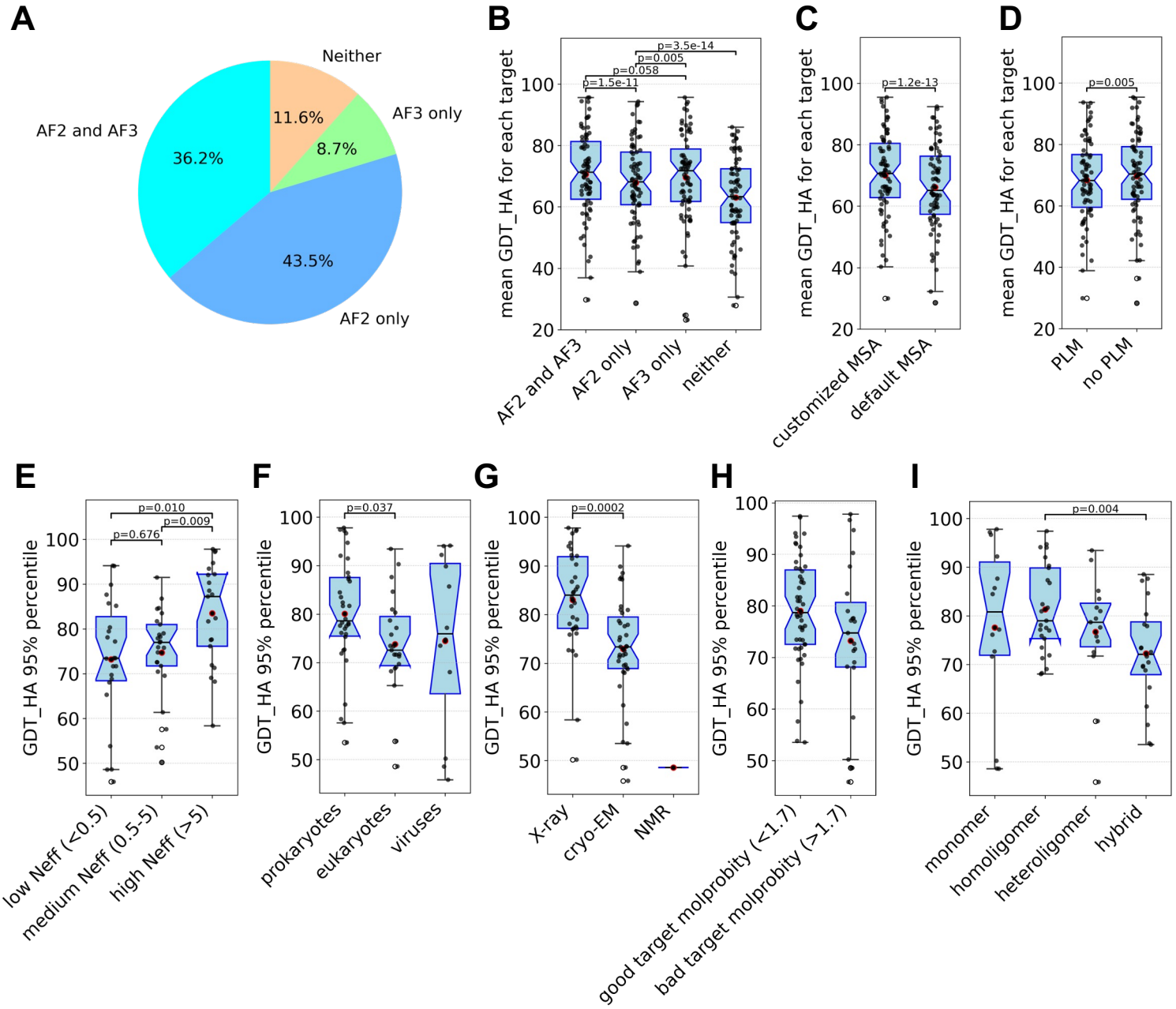
79. Antczak, P. L. M., Ratajczak, T., Lukasiak, P. & Blazewicz, J. SphereGrinder - reference structure-based tool for quality assessment of protein structural models. in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 665–668 (2015). doi:10.1109/BIBM.2015.7359765.

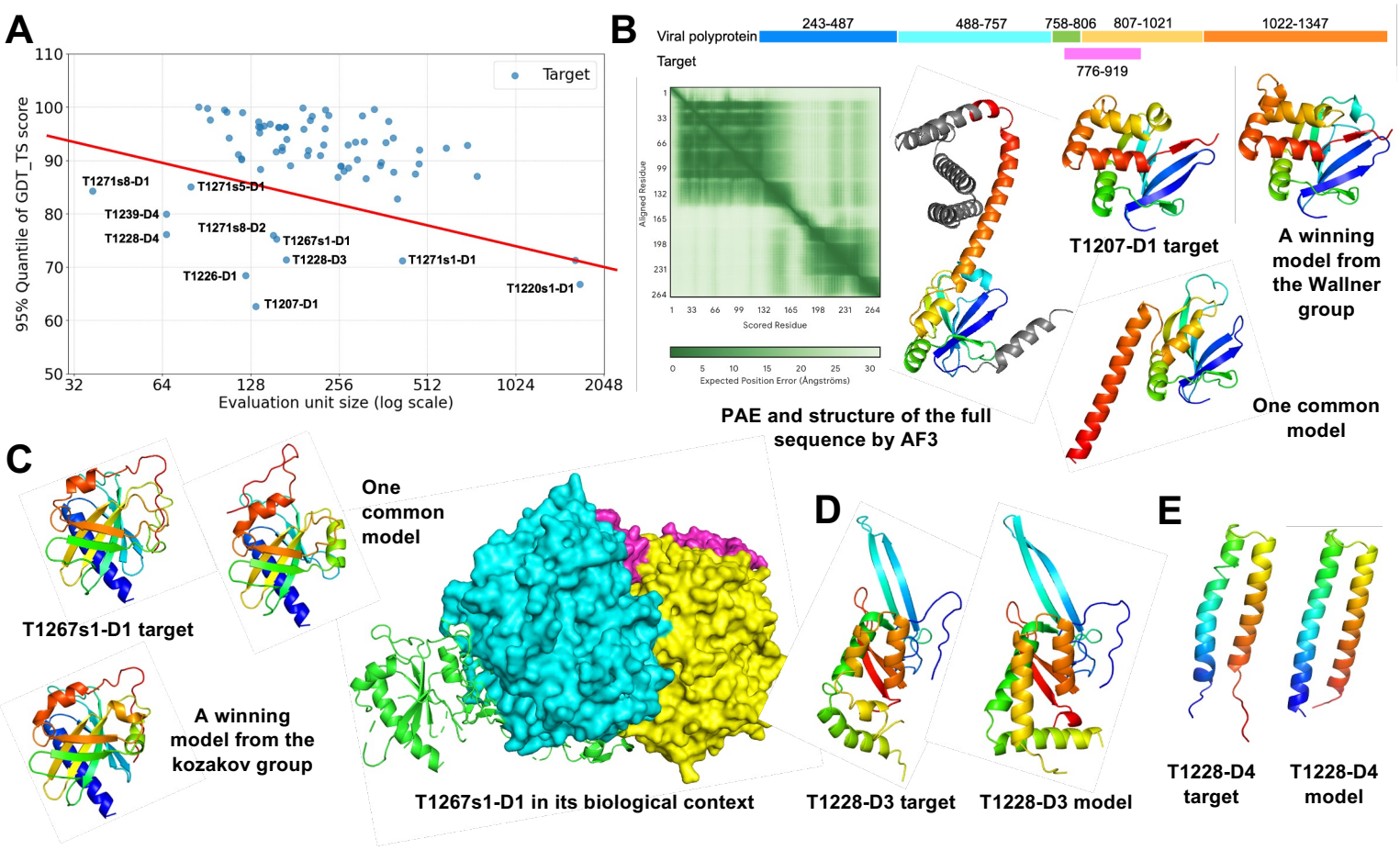
80. Olechnovič, K., Kulberkytė, E. & Venclovas, Č. CAD-score: A new contact area difference-based function for evaluation of protein structural models. *Proteins Struct. Funct. Bioinforma.* **81**, 149–162 (2013).

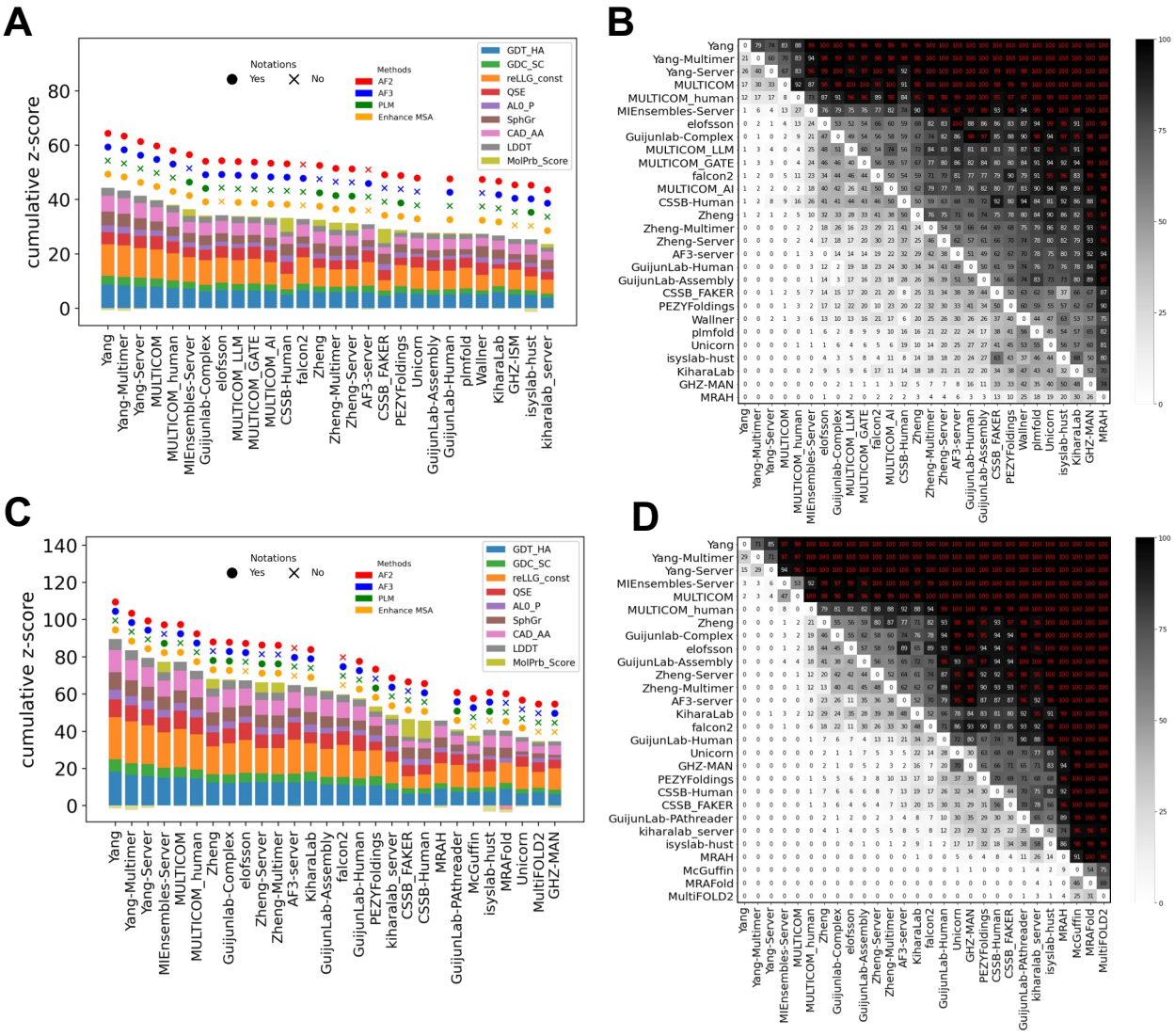
81. Mariani, V., Kiefer, F., Schmidt, T., Haas, J. & Schwede, T. Assessment of template based protein structure predictions in CASP9. *Proteins Struct. Funct. Bioinforma.* **79**, 37–58 (2011).

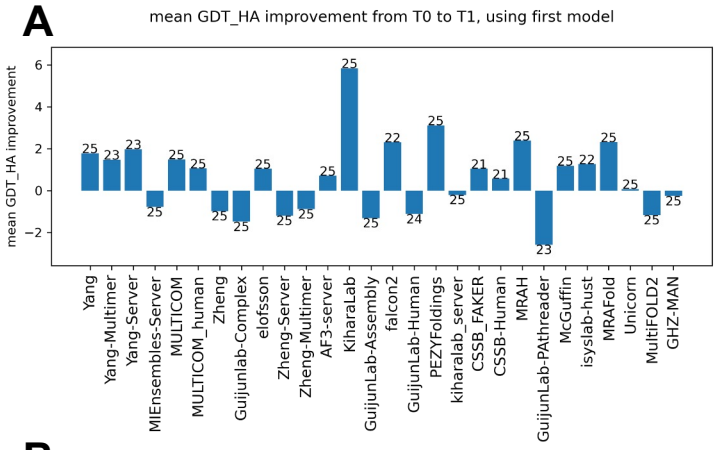
82. Biggs, J. factor_analyzer Documentation.

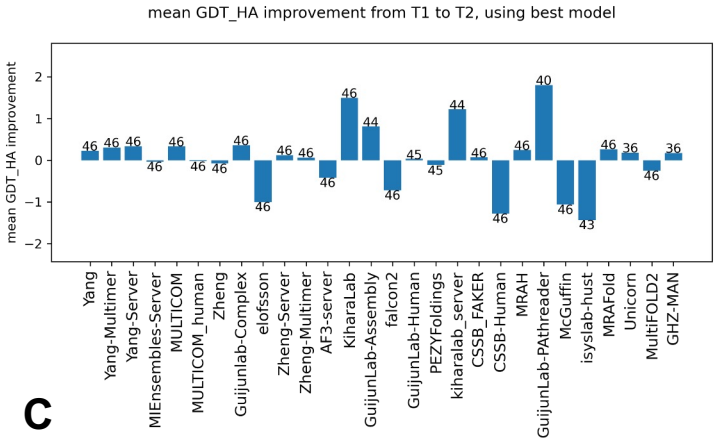

Group strategies
AF2
Y
N
N/A
AF3
Y
N
N/A
PLM
Y
N
N/A
MSA
Enhanced MSA
Use MSA
No MSA
N/A
GDT_HA scale
80
60
40
20
① ② ③ ④ ⑤
AF2
AF3
PLM
MSA
Target features
① taxonomy
prokaryotes (33)
low eukaryotes (12)
high eukaryotes (11)
viral (12)
singleton (2)
transposon (1)
nanobody (2)
archaea (1)
② target source
mono (14)
homo (25)
hetero (15)
hybrid (20)
③ experiment
X-ray (34)
cryo-EM (39)
NMR (1)
④ difficulty
easy (24)
medium (41)
hard (9)
⑤ target size
small (21)
large (53)

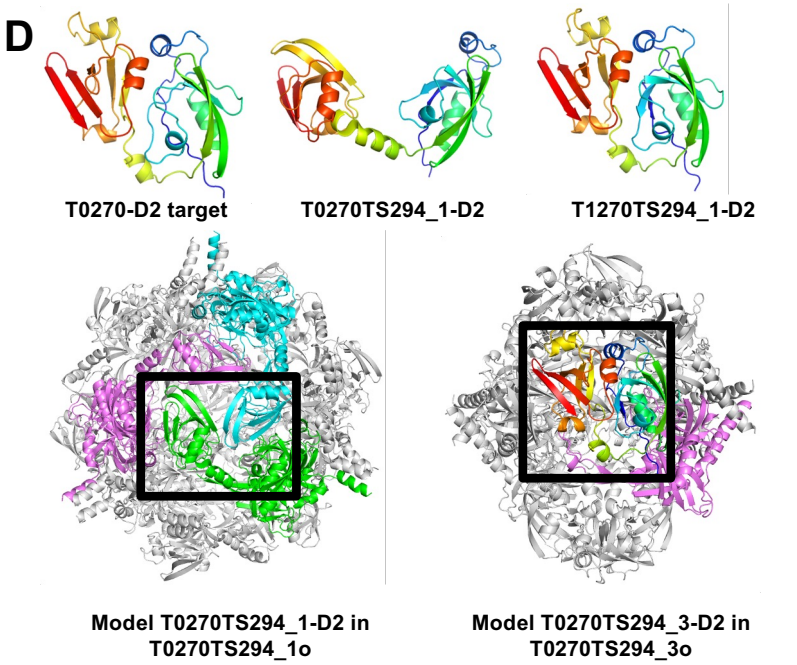

A
Neither
AF3 only
AF2 and AF3
AF2 only
11.6%
8.7%
36.2%
43.5%
B
mean GDT_HA for each target
p=1.5e-11
p=0.058
p=0.005
p=3.5e-14
AF2 and AF3
AF2 only
AF3 only
neither
C
mean GDT_HA for each target
p=1.2e-13
customized MSA
default MSA
D
mean GDT_HA for each target
p=0.005
PLM
no PLM
E
GDT_HA 95% percentile
p=0.676
p=0.009
p=0.010
low Neff (<0.5)
medium Neff (0.5-5)
high Neff (>5)
F
GDT_HA 95% percentile
p=0.037
prokaryotes
eukaryotes
viruses
G
GDT_HA 95% percentile
p=0.0002
X-ray
cryo-EM
NMR
H
GDT_HA 95% percentile
good target molprobity (<1.7)
bad target molprobity (>1.7)
I
GDT_HA 95% percentile
p=0.004
monomer
homoligomer
heteroligomer
hybrid

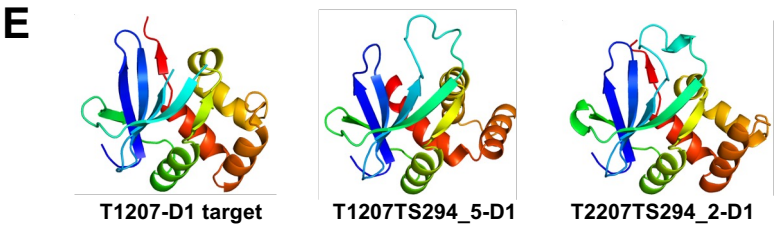

A
Target
95% Quantile of GDT_TS score
Evaluation unit size (log scale)
T1271s8-D1
T1271s5-D1
T1239-D4
T1228-D4
T1271s8-D2
T1267s1-D1
T1228-D3
T1271s1-D1
T1226-D1
T1207-D1
T1220s1-D1
B
Viral polyprotein
243-487
488-757
758-806
807-1021
1022-1347
Target
776-919
Aligned Residue
Scored Residue
Expected Position Error (Ångströms)
PAE and structure of the full sequence by AF3
T1207-D1 target
A winning model from the Wallner group
One common model
C
T1267s1-D1 target
One common model
A winning model from the kozakov group
T1267s1-D1 in its biological context
D
T1228-D3 target
T1228-D3 model
E
T1228-D4 target
T1228-D4 model

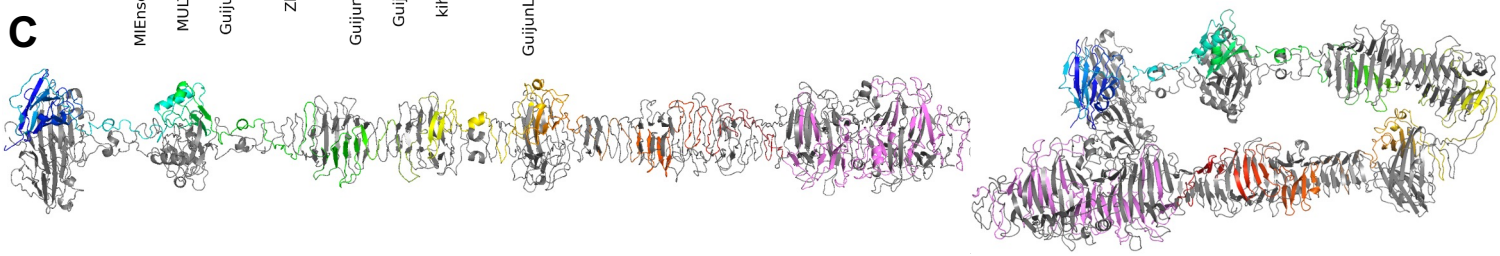

A
B
C
D
cumulative z-score
Notations
Yes
No
Methods
AF2
AF3
PLM
Enhance MSA
GDT_HA
GDC_SC
reLLG_const
QSE
AL0_P
SphGr
CAD_AA
LDDT
MolPrb_Score

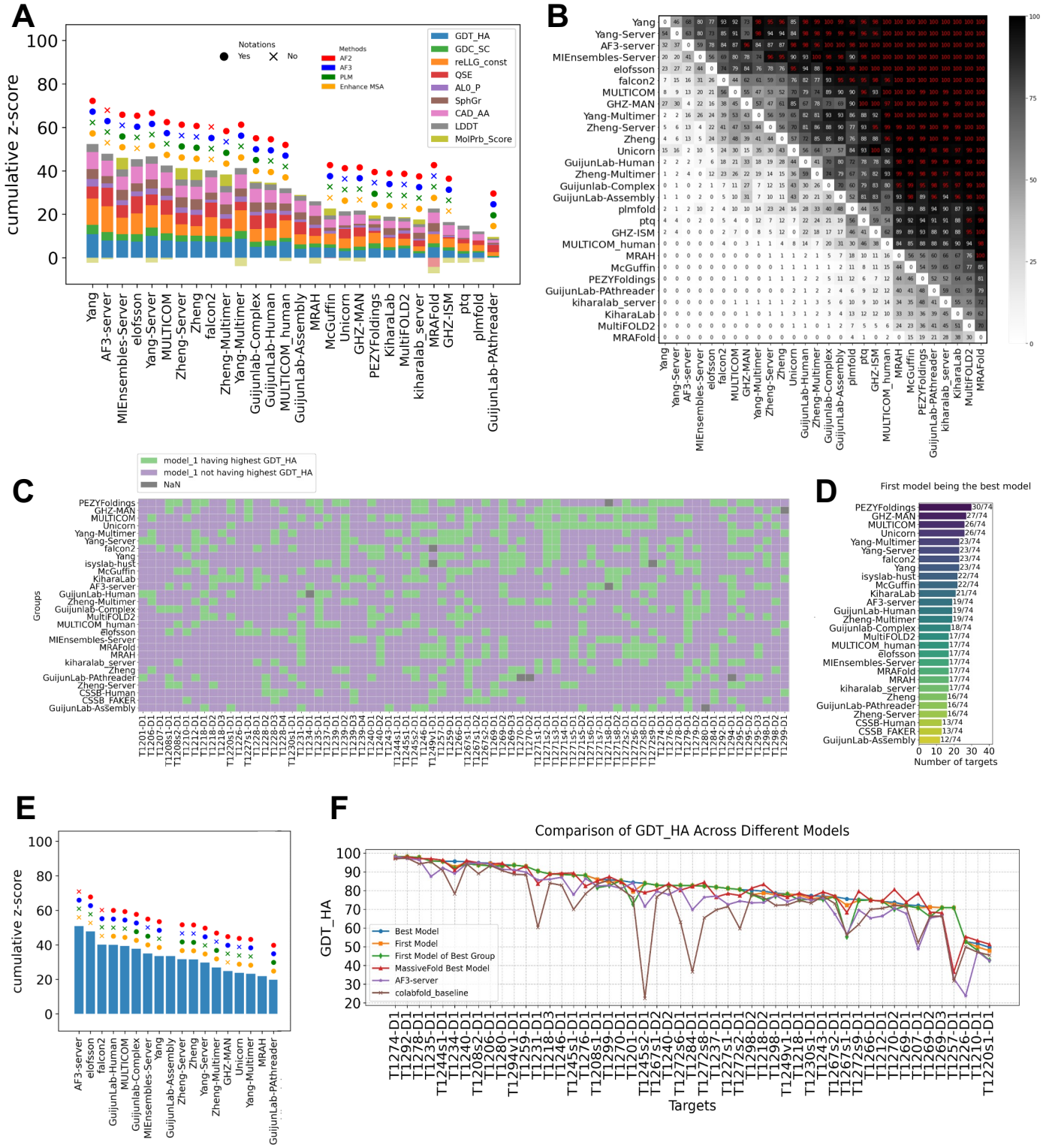

A
mean GDT_HA improvement from T0 to T1, using first model
mean GDT_HA improvement
Yang
Yang-Multimer
Yang-Server
MIEnsembles-Server
MULTICOM
MULTICOM_human
Zheng
Guijunlab-Complex
elofsson
Zheng-Server
Zheng-Multimer
AF3-server
KiharaLab
GuijunLab-Assembly
falcon2
GuijunLab-Human
PEZYFoldings
kiharalab_server
CSSB_FAKER
CSSB-Human
MRAH
GuijunLab-PAthreader
McGuffin
isyslab-hust
MRAFold
Unicorn
MultiFOLD2
GHZ-MAN

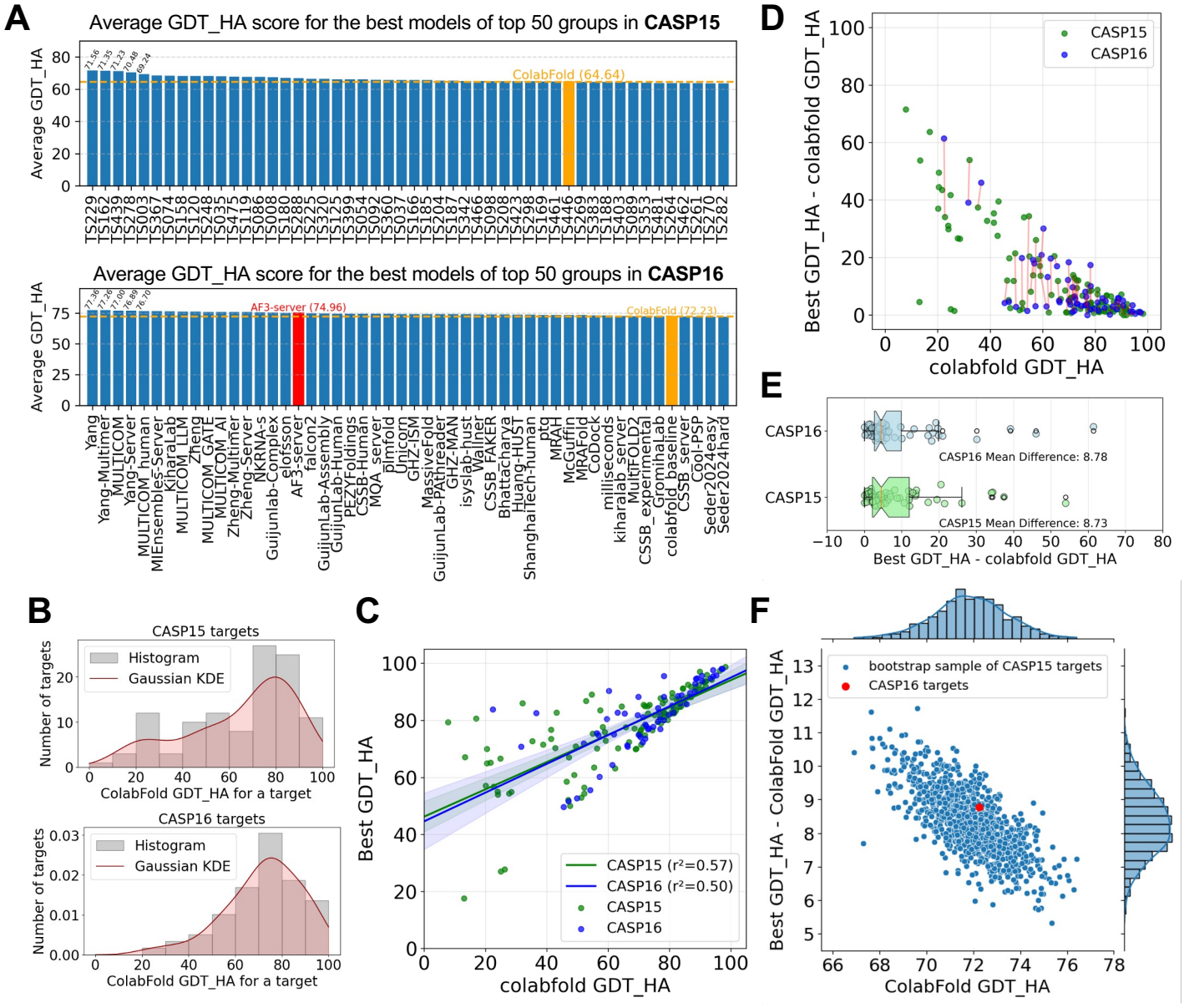

B
mean GDT_HA improvement from T1 to T2, using best model
mean GDT_HA improvement
Yang
Yang-Multimer
Yang-Server
MIEnsembles-Server
MULTICOM
MULTICOM_human
Zheng
Guijunlab-Complex
elofsson
Zheng-Server
Zheng-Multimer
AF3-server
KiharaLab
GuijunLab-Assembly
falcon2
GuijunLab-Human
PEZYFoldings
kiharalab_server
CSSB_FAKER
CSSB-Human
MRAH
GuijunLab-PAthreader
McGuffin
isyslab-hust
MRAFold
Unicorn
MultiFOLD2
GHZ-MAN

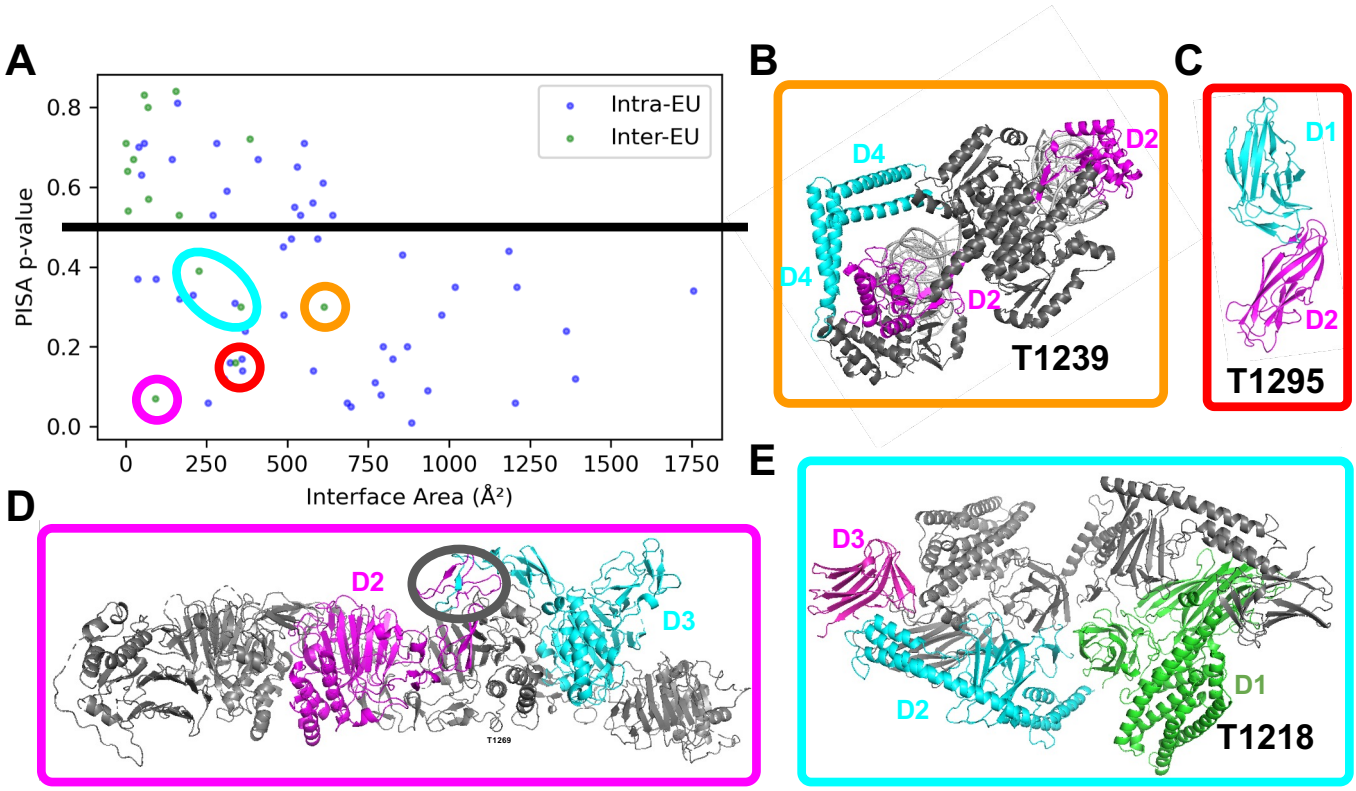

D
T0270-D2 target
T0270TS294_1-D2
T1270TS294_1-D2
Model T0270TS294_1-D2 in T0270TS294_1o
Model T0270TS294_3-D2 in T0270TS294_3o


E
T1207-D1 target
T1207TS294_5-D1
T2207TS294_2-D1


C
Model T1257TS345_1-D1 in its predicted trimmer context
Model T0257TS345_1-D1 in its predicted trimmer context


A
cumulative z-score
Notations
Yes
No
Methods
AF2
AF3
PLM
Enhance MSA
GDT_HA
GDC_SC
reLLG_const
QSE
AL0_P
SphGr
CAD_AA
LDDT
MolPrb_Score
Yang
AF3-server
MIEnsembles-Server
elofsson
Yang-Server
MULTICOM
Zheng-Server
Zheng
falcon2
Zheng-Multimer
Yang-Multimer
Guijunlab-Complex
GuijunLab-Human
MULTICOM_human
GuijunLab-Assembly
MRAH
McGuffin
Unicorn
GHZ-MAN
PEZYFoldings
KiharaLab
MultiFOLD2
kiharalab_server
MRAFold
GHZ-ISM
ptq
plmfold
GuijunLab-PAthreader
B
C
model_1 having highest GDT_HA
model_1 not having highest GDT_HA
NaN
Groups
D
First model being the best model
Number of targets
E
cumulative z-score
F
Comparison of GDT_HA Across Different Models
GDT_HA
Best Model
First Model
First Model of Best Group
MassiveFold Best Model
AF3-server
colabfold_baseline
Targets


A
Average GDT_HA score for the best models of top 50 groups in CASP15
Average GDT_HA
ColabFold (64.64)
Average GDT_HA score for the best models of top 50 groups in CASP16
AF3-server (74.96)
ColabFold (72.23)
B
CASP15 targets
Histogram
Gaussian KDE
Number of targets
ColabFold GDT_HA for a target
CASP16 targets
C
Best GDT_HA
colabfold GDT_HA
CASP15 (r²=0.57)
CASP16 (r²=0.50)
CASP15
CASP16
D
Best GDT_HA - colabfold GDT_HA
colabfold GDT_HA
E
CASP16
CASP15
CASP16 Mean Difference: 8.78
CASP15 Mean Difference: 8.73
Best GDT_HA - colabfold GDT_HA
F
bootstrap sample of CASP15 targets
CASP16 targets
Best GDT_HA - ColabFold GDT_HA
ColabFold GDT_HA


A
Intra-EU
Inter-EU
PISA p-value
0.0
0.2
0.4
0.6
0.8
0
250
500
750
1000
1250
1500
1750
Interface Area (Å²)
B
D4
D2
D4
D2
T1239
C
D1
D2
T1295
D
D2
D3
T1269
E
D3
D2
D1
T1218