Critical Assessment of Methods of Protein Structure Prediction (CASP) – Round XV

Andriy Kryshtafovych[1], Torsten Schwede[2], Maya Topf[3], Krzysztof Fidelis[1], and John Moult[4]

[1]University of California Davis Genome Center
[2]Universitat Basel Department Biozentrum
[3]Universitatsklinikum Hamburg-Eppendorf Institut fur Experimentelle Pharmakologie und Toxikologie
[4]Institute for Bioscience and Biotechnology Research Shady Grove

October 6, 2023

Abstract

Computing protein structure from amino acid sequence information has been a long-standing grand challenge. CASP (Critical Assessment of Structure Prediction) conducts community experiments aimed at advancing solutions to this and related problems. Experiments are conducted every two years. The 2020 experiment (CASP14) saw major progress, with the second generation of deep learning methods delivering accuracy comparable with experiment for many single proteins. There is an expectation that these methods will have much wider application in computational structural biology. Here we summarize results from the most recent experiment, CASP15, in 2022, with an emphasis on new deep learning-driven progress. Other papers in this special issue of Proteins provide more detailed analysis. For single protein structures, the AlphaFold2 deep learning method is still superior to other approaches, but there are two points of note. First, although AlphaFold2 was the core of all the most successful methods, there was a wide variety of implementation and combination with other methods. Second, using the standard AlphaFold2 protocol and default parameters only produces the highest quality result for about two thirds of the targets, and more extensive sampling is required for the others. The major advance in this CASP is the enormous increase in the accuracy of computed protein complexes, achieved by the use of deep learning methods, although overall these do not fully match the performance for single proteins. Here too, AlphaFold2 based method perform best, and again more extensive sampling than the defaults is often required. Also of note are the encouraging early results on the use of deep learning to compute ensembles of macromolecular structures. Critically for the usability of computed structures, for both single proteins and protein complexes, deep learning derived estimates of both local and global accuracy are of high quality, however the estimates in interface regions are slightly less reliable. CASP15 also included computation of RNA structures for the first time. Here, the classical approaches produced better agreement with experiment than the new deep learning ones, and accuracy is limited. Also, for the first time, CASP included the computation of protein-ligand complexes, an area of special interest for drug design. Here too, classical methods were still superior to deep learning ones. Many new approaches were discussed at the CASP conference, and it is clear methods will continue to advance.

Critical Assessment of Methods of Protein Structure Prediction (CASP) –

Round XV

Running title: CASP15 Overview

Andriy Kryshtafovych[1], Torsten Schwede[2], Maya Topf[3], Krzysztof Fidelis[1] & John Moult[4*]

[1]Genome Center, University of California, Davis, CA, USA

[2]University of Basel, Biozentrum & SIB Swiss Institute of Bioinformatics, Basel, Switzerland

[3]Centre for Structural Systems Biology, Leibniz-Institut fur Experimentelle Virologie and Universitätsklinikum Hamburg-Eppendorf (UKE), Hamburg, Germany.

[4]Institute for Bioscience and Biotechnology Research, Rockville, MD, USA, and Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD, USA

*Corresponding author:

John Moult

tel: 240-314 6241

email: jmoult@umd.edu

Keywords: Protein Structure Prediction, Community Wide Experiment, CASP

ABSTRACT

Computing protein structure from amino acid sequence information has been a long-standing grand challenge. CASP (Critical Assessment of Structure Prediction) conducts community experiments aimed at advancing solutions to this and related problems. Experiments are conducted every two years. The 2020 experiment (CASP14) saw major progress, with the second generation of deep learning methods delivering accuracy comparable with experiment for many single proteins. There is an expectation that these methods will have much wider application in computational structural biology. Here we summarize results from the most recent experiment, CASP15, in 2022, with an emphasis on new deep learning-driven progress. Other papers in this special issue of Proteins provide more detailed analysis. For single protein structures, the AlphaFold2 deep learning method is still superior to other approaches, but there are two points of note. First, although AlphaFold2 was the core of all the most successful methods, there was a wide variety of implementation and combination with other methods. Second, using the standard AlphaFold2 protocol and default parameters only produces the highest quality result for about two thirds of the targets, and more extensive sampling is required for the others. The major advance in this CASP is the enormous increase in the accuracy of computed protein complexes, achieved by the use of deep learning methods, although overall these do not fully match the performance for single proteins. Here too, AlphaFold2 based method perform best, and again more extensive sampling than the defaults is often required. Also of note are the encouraging early results on the use of deep learning to compute ensembles of macromolecular structures. Critically for the usability of computed structures, for both single proteins and protein complexes, deep learning derived estimates of both local and global accuracy are of high quality, however the estimates in interface regions are slightly less reliable. CASP15 also included computation of RNA structures for the first time. Here, the classical approaches produced better agreement with experiment than the new deep learning ones, and accuracy is limited. Also, for the first time, CASP included the computation of protein-ligand complexes, an area of special interest for drug design. Here too, classical methods were still superior to deep learning ones. Many new approaches were discussed at the CASP conference, and it is clear methods will continue to advance.

1 | INTRODUCTION

Computing the three-dimensional structure of protein molecules from their amino acid sequence first emerged as an aspiration in the 1960s (1), and since then many different approaches have been tried. CASP (Critical Assessment of Structure Prediction) was introduced in 1994 with the aim of accelerating progress by rigorously assessing the performance of methods through a community-wide experiment. Every two years, members of the experimental community are asked to provide information about soon-to-be-released structures, and the amino acid sequence information is passed on to the computational community, with the challenge of calculating the corresponding three-dimensional atomic structures. The similarity between computed and experimental structures is then examined by independent assessors with the support of the UC Davis Center for CASP (https://predictioncenter.org), the outcome discussed at an international meeting, and findings published in a special journal issue. This paper summarizes the results of the 15th experiment,

held in 2022, and the authors are the experiment organizers. Other papers in this issue of PROTEINS are by the assessors and research groups with leading performances in various aspects of the experiment. Full details of the experiment, including targets ad results are at (https://predictioncenter.org). CASP is complemented by CAMEO (2), a continuous evaluation of computed structure accuracy that utilizes PDB weekly releases as targets.

CASP has seen massive progress in model accuracy over the course of the experiments. Initially, through homology modeling methods. But until recently, there was very limited effectiveness for structures where homology was not applicable, and accuracy very rarely approached that of experimental methods. In CASP13 (2018), that began to change dramatically through the effective application of convolutional neural networks. That approach resulted in models with correct folds for the majority of targets (3). That was followed by the introduction of attention-based networks and other algorithmic advances in CASP14 (2020), resulting in accuracy judged to be competitive with experiment for about two-thirds of the single protein targets (4). Although multiple participating research groups made progress in 2020, by far the most accurate results were obtained with AlphaFold2 (AF2) from the company DeepMind.

CASP15 (2022) builds on these earlier results. Although agreement with experiment for single protein structures had largely converged by 2020, key questions remained. These include whether observed limitations, particularly for shallow sequence alignments, would be overcome, how different protocols built around AF2 would perform, and whether other new methods would match or exceed AF2 performance. Of great interest was whether and to what extent deep learning methods would prove effective in addressing other problems in computational structural biology. The scope of CASP was to allow fuller investigation of that. One of the areas of most interest for the application of deep learning is that of protein complexes. For the last five rounds, CASP has included that category, in collaboration with CAPRI (5). For the first time, this CASP also includes categories on calculating RNA structure from sequence (in collaboration with RNA Puzzles (6)) and of calculating the structure of protein-small ligand complexes, particularly relevant to drug design. These are both areas where papers have suggested deep learning may make a major difference (7, 8). CASP also continued its long-standing category on methods to estimate the accuracy of models with new emphasis on the estimated accuracy of protein complexes, a critical factor for structure usefulness. This CASP also includes a new category for modeling ensembles of macromolecular conformations (9).The original framing of the protein folding problem was done 50 years ago (1), when there were only experimental structures for a few small, simple, highly ordered proteins. We now appreciate that proteins and RNA may adopt different conformations, both under the same conditions and in response to changes such as ligand binding or mutations so that speaking of 'the conformation' can become meaningless. Thus, computational methods should be able to reproduce multiple observed structural states. Results for all these categories are summarized in this paper; other papers in this Proteins issue describe detailed assessment results (9-14) and comments on the outcome by some of those contributing targets (15, 16). The issue also contains papers by selected participating groups. Some categories no longer seen as relevant were dropped: contact prediction which is now an integral part of deep learning methods, and refinement of initial models. The latter category, although partly successful in earlier CASPs, in 2020 could not produce models of comparable accuracy to those obtained with AF2. The 'data assisted' and 'function analysis' categories, although still relevant, were also not included in this CASP.

As set out below, this was another very exciting CASP round, with areas of major progress. There are remaining major limitations in some areas, but in all there are clear prospects for further progress.

2 | RESULTS

2.1 | Scope of the experiment

In the Spring and Summer of 2022, CASP solicited 103 yet-to-be-published structures as potential modeling targets for the CASP15 experiment. The structures were provided by 48 structure determination groups from 14 countries all over the world. They include single-sequence protein molecules, protein-protein complexes, RNA molecules, RNA-protein complexes, and protein-ligand complexes. Ninety-eight proposed targets we-

re deemed suitable for the CASP experiment and were gradually released for prediction over a period of three months (May 2 through July 29, 2022). Together with target sequences, predictors were provided the information on the oligomerization state of the targets and availability of bound organic ligands or ions (17). All in all, CASP15 offered 47 multimeric protein complexes for modeling in the Protein assembly category (40 assessed), 81 monomeric proteins or subunits of multimeric complexes in the Tertiary Structure category (77 assessed, divided into 112 Evaluation units (18)), 12 RNA structures in the RNA category, 23 protein-ligand complexes in the Ligand binding category, and two protein-RNA complexes. 89 research labs from 17 countries submitted models in one or more categories, with the USA (30 labs) and China (29 labs) most represented. Some labs created more than one participating group, so that 162 groups were assessed altogether. Of these, 56 are server groups, where target sequences are sent directly to a CASP-registered computer and a response is required within 72 hours. For non-server groups, usually there was a three-week window in which to submit models, allowing time for human intervention.

2.2 | Three-dimensional protein structure

As noted earlier, the 2020 CASP round (CASP14) saw a dramatic advance in the accuracy of computational models of single proteins, with the accuracy of many models rivaling that of experiment (4). The few major failures were for oligomeric proteins (for example viral coat components) considered in isolation and some targets with very shallow multiple sequence alignment. Agreement with experiment also tended be lower for the few NMR targets.

For CASP15, interest centered around whether groups other than DeepMind could approach experimental accuracy, whether new approaches would rival AlphaFold2, and whether limitations for shallow sequence alignments would be overcome. DeepMind did not participate in this CASP, but the Elofsson group ran each target through the local standard installation of Alphafold2, providing a baseline.

Figure 1 shows one of the many examples of a protein modeled to high accuracy in this CASP.

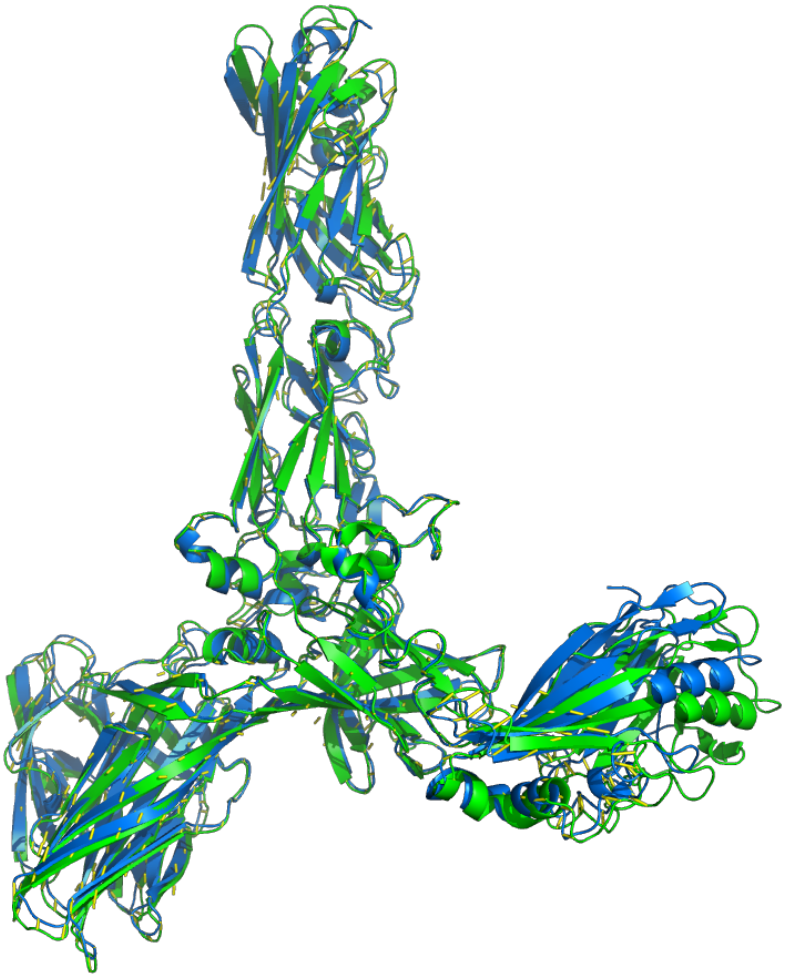
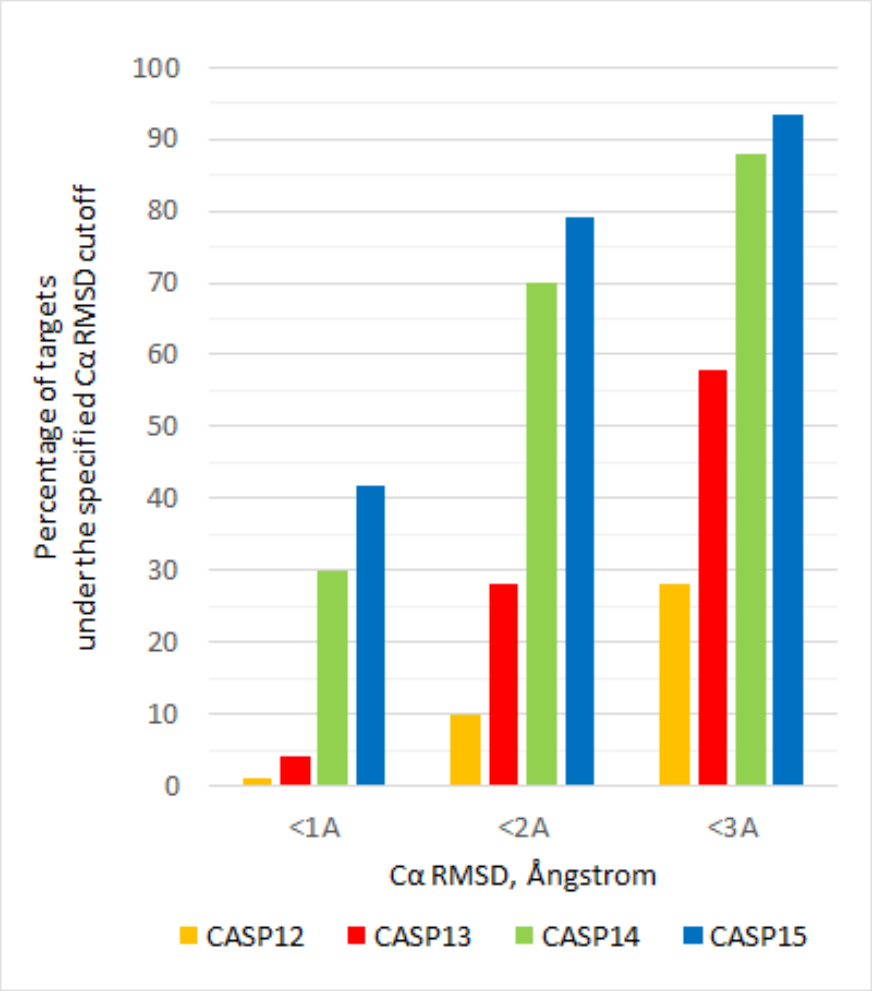


Figure 1: *Superposition of a large protein target, T1154, a 1040 residue archaeal S-layer protein (green) and the closest calculated structure (blue). Misalignment of the lower right-hand domain likely reflects interdomain flexibility. For the assessment, the structure was divided into two evaluation units: the lower right-hand domain, and the rest of the structure. Corresponding highest GDT_TS values are 88 and 90%. There are no detectable templates for either unit.*

Figure 2 shows performance in terms of the agreement between the best CASP models and the corresponding experimental structures. The Cα RMSD is below 3Å for over 90% of CASP15 targets, and less than 1Å for 40% of the targets. These are impressive results, even though only modest increases over the CASP14 (2020) performance.



Φιγυρε 2: *Περσενταγε οφ ταργετς μοδελεδ το τηρεε â RMΣΔ αςςυραςψ τηρεσηολδς (1, 2 ανδ 3Å) ιν τηε φουρ μοστ ρεςεντ ΆΣΠ εξπεριμεντς. Τηε φραςτιον οφ ηιγη αςςυραςψ στρυςτυρες ινςρεασεδ δραματιςαλλψ φρομ 2016 (ΆΣΠ12) το 2018 (ΆΣΠ13) βεςαυσε οφ τηε ιντροδυςτιον οφ εφφεςτιε δεεπ λεαρνινγ μετηοδς ανδ αγαιν φρομ 2018 το 2020 (ΆΣΠ14) ωιτη τηε ιντροδυςτιον οφ τηε ΑλπηαΦολδ2 δεεπ λεαρνινγ μετηοδ. Ινςρεασες φρομ 2020 το 2022 (ΆΣΠ15) αρε μορε μοδεστ λικελψ βεςαυσε ιν ΆΣΠ14 μανψ ςομπυτεδ στρυςτυρες ωερε αλρεαδψ ωιτηιν εξπεριμενταλ υνςερταιντψ, σο τηερε ις νοτ μυςη ροομ φορ φυρτηερ ιμπροεμεντ. RMΣΔ (ροοτ μεαν σχυαρε δειατιον) ινςλυδες αλλ ςομμον ρεσιδυες ιν τηε εξπεριμενταλ ανδ ςομπυτεδ στρυςτυρες.*

Figure 3 shows progress over all the CASP experiments using the more robust GDT_TS measure of backbone agreement, and is an update of equivalent figures shown in earlier CASP overview papers. As in the previous 2020 CASP14 (blue line), and consistent with the RMSD result in figure 2, the majority of best models (black line and open circles) approach experimental accuracy (approximately 90% on the GDT_TS scale), and overall best performance is very similar. There is only a small fall-off with the extent to which homology modeling can be utilized (X axis difficulty scale). This result is consistent with expectation, since once experimental accuracy is reached there is no way of measuring further improvement. Of note, best server performance (dotted line) is similar to that returned by groups where human intervention was possible, showing that most procedures could be fully automated.

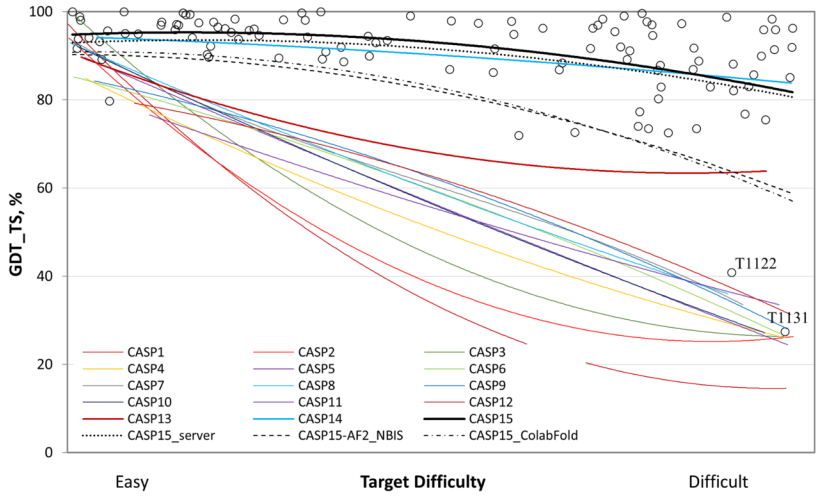


Figure 3: *CASP15 performance for protein tertiary structure, compared with earlier CASPs. The Y axis shows backbone agreement with experiment in GDT_TS units (19). On this scale, a random model scores approximately 20 to 30%, a correctly folded model around 50, and a model within experimental accuracy, around 90. Open circles show CASP15 best model results for each target, trend lines show performance for each CASP. Overall best performance of CASP15 (black line) is similar to that in CASP14 (blue line). The dotted line shows best server performance in CASP15. AlphaFold2-based methods dominate best performance. However, performance with standard AlphaFold2 protocols available at the time of the experiment is lower (dashed black lines). The X-axis difficulty scale represents the extent to which homology-based could be utilized. Targets in each CASP are ordered by difficulty calculated as a cumulative rank of the sequence identity and the coverage of the target by the best homologous structure available at the time of each experiment. The templates are found by running Foldseek (20) and LGA (21) versus experimental structures deposited to the PDB.*

All the most successful CASP15 methods are based on the CASP14 DeepMind AlphaFold2 method, AF2, and subsequent DeepMind updates. But strikingly, two versions of standard AF2 procedures (dashed black lines), one models from the ColabFold server (22) and one using a local installation of AF2 with default parameters (23) have substantially worse performance, and performance does fall off markedly as homology information weakens. We consulted DeepMind about this finding, and they undertook to run the CASP targets internally. The results from that process are not shown since they are not official CASP outcomes, but overall are very similar to the CASP15 best performance line. The primary reason for the difference in best CASP15 and standard AF2 performance appears to be that the most successful methods all used greater sampling of possible structures than the AF2 defaults, including different combinations of a larger number of seeds, more recycles, and network dropout. Most also used customized and sometimes enhanced multiple sequence alignments.

There are two important conclusions. First, as of mid-2022, AF2 based methods were clearly more accurate than others. The next best performance was from RosettaFold (24), a method developed by the Baker group following AF2 principles (note this is RosettaFold version 1; version 2 has since been released (25)). Participating deep learning Large Language Models (LLMs) did not perform well. Second, to get the best results from AF2 it is often necessary to sample more extensively than the default parameters allow and to carefully choose and adjust the multiple sequence alignment. For about 1/3 of the targets, there is a gain in GDT_TS of 10% or more in using the improved methods with enhanced sampling. For a few targets the difference in GDT_TS is more than two-fold, but nearly all of those cases are domains of large proteins. As

discussed below, these targets are generally more challenging.

Figure 3 shows that while many targets achieve a GDT_TS score of at least 90%, there are also some low scoring targets. The lowest scoring target is T1131. This protein is a member of a rapidly evolving family of proteins in aphids, likely involved in creating a gall on the plant host (26). There were no detectable sequence relatives outside that family, and the family sequences were only available on a specialized web site. Thus, this appears to be a clear example of single sequence being insufficient to produce a good model, a deficiency also seen in DeepMind's benchmarking (27) and in the previous CASP (4). It was hoped that Large Language Models would remove this limitation (for example (28)), but in CASP15 that was not yet the case. T1122, the next lowest performing target, has a best GDT_TS of just over 40%. This protein also has a shallow sequence alignment, but the crystal structure may be highly influenced by the very tight intermolecular interactions (21% crystal solvent content). Supplementary Figure 1 shows that in CASP15 there is a tendency for performance to fall off with decreasing alignment depth, and this is much more pronounced when the ratio of alignment depth to target length is below 0.1 (less than 10 appropriately diverse sequences per 100 residues.) As in the previous CASP, there are also targets with shallow alignments which nevertheless have high agreement with experiment. That is, consistent with DeepMind's benchmarking (27), shallow alignments sometimes result in poor models but in other instances are fine. There are nine other targets with a best GDT_TS score between 70 and 80. One of these is an NMR target (T1155) containing a large flexible loop (personal communication, Luciano Abriata), so it may have multiple conformations. As discussed later, the new CASP ensembles category is beginning to shed light on this type of issue. The others are all domains of large targets (˜1200 to over 4000 residues), and in general, performance seems to be a little worse for domains of very large proteins. Unlike in the previous CASP, there is no fall-off in performance with experimental resolution.

2.2.1 | Domain-domain interactions

For CASP assessment, some protein structures are divided into multiple evaluation units, splitting along domain boundaries, on the basis of relatively low GDT_TS values for the whole structure compared to the separate EUs (18). In this sense, apparently high accuracy may apply to only part of a structure, giving a misleading impression. To investigate this point and other aspects of domains, the protein assembly assessor analyzed the accuracy of the relevant domain-domain interfaces using the same methods as for subunit-subunit interfaces (11). In CASP15, 22 of the 77 tertiary structure targets were split into two or more EUs resulting in 112 evaluation units (18). Of these 22 structures, 15 have insufficient interactions between the EUs to stabilize a fixed relationship (using a threshold of less than 10 interface inter-residue contacts, where a contact is any residue-residue interatomic distance of less than 5Å). The remaining seven targets have 21 inter-EU interactions in total. Domain-domain interface agreement with experiment is expressed as the Interface Contact Score (ICS), equivalent to F1 for residue-residue contacts across an interface (maximum value 1.0). Poor ICS scores tend to occur in larger proteins. Of the 21 interfaces, only two are in proteins of less than 1000 residues and these have relatively high ICS values of 0.88 and 0.99. There is generally more flexibility across domain interfaces than within domains, and ICS is sensitive to small changes. Assuming ICS values greater than 0.8 approach experimental uncertainty limits, all interfaces in shorter proteins and seven of the 19 interfaces in larger proteins (>1000 residues) are accurate.

2.3 | Structure of protein assemblies

Previous benchmarking studies, for example (29), had suggested that deep learning methods, particularly AlphaFold2, would also be effective for protein assemblies, and so this category is of especial interest. The analysis was conducted in close collaboration with our colleagues at CAPRI (Critical Assessment of PRedicted Interactions (5)). Two assessments were performed, one by CASP (11) and one by CAPRI (30), providing two views. Here we summarize the main conclusions from the CASP15 assessment.

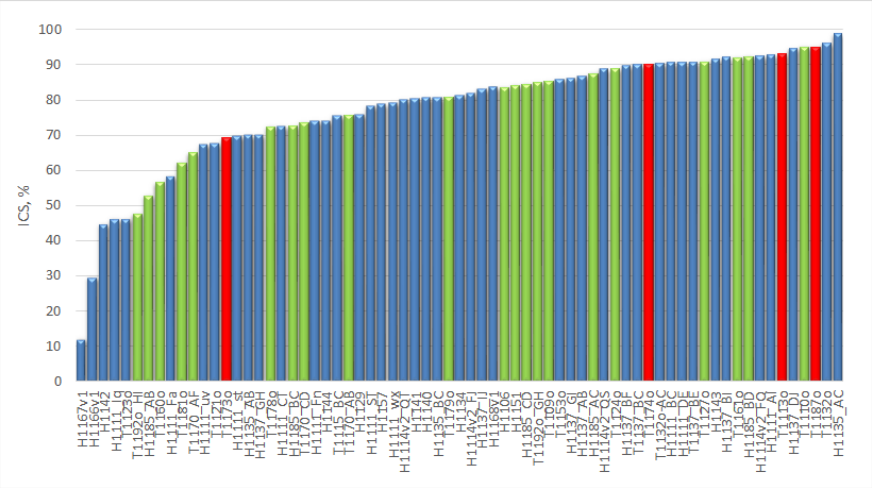
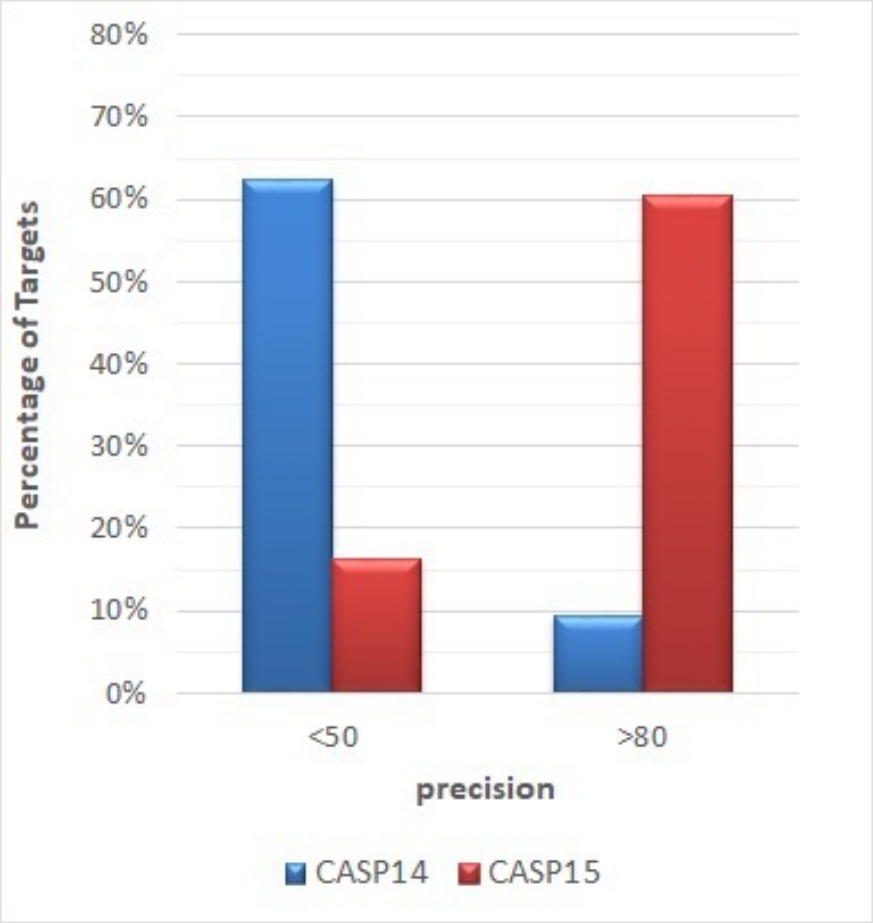


Figure 4: *CASP15 Protein assembly subunit interface agreement with experiment. Bars show the best agreement obtained between computed and experimental interfaces, using the Interface Contact Score (ICS). Green bars indicate interfaces in targets where there are homologous structures for all component subunits and interfaces. Blue bars are those where there are homologous structures for some subunits and interfaces, and red those where there is no structural homology. More than half the interfaces reach a score of 0.8 or higher, and so may approach experimental accuracy. About a quarter reach a more stringent criterion of ICS > 90.*

In CASP15 a total of 40 protein assembly targets were assessed, ranging from simple obligate dimers to very large complexes with up to 11 polypeptide chains and up to more than 4000 amino acids. Figure 4 shows the distribution of interface structure agreement with experiment in terms of the Interface Contact score (ICS), the F1 measure of contact agreement introduced earlier. Examination of interfaces with lowest ICS values (<0.6) is informative. Three of the eight targets bearing the poorly modeled interfaces are immune complexes, discussed below. Four interfaces are from large complexes, which tend to be a slightly lower accuracy overall, and one is for a viral coat protein homodimer (T1123). The multiple sequence alignment for the monomer of this homo-dimeric target is shallow (10 sequences), but the subunit's best models are of high accuracy (best GDT_TS of 87), so it is unclear why this interface has poor agreement with experiment.



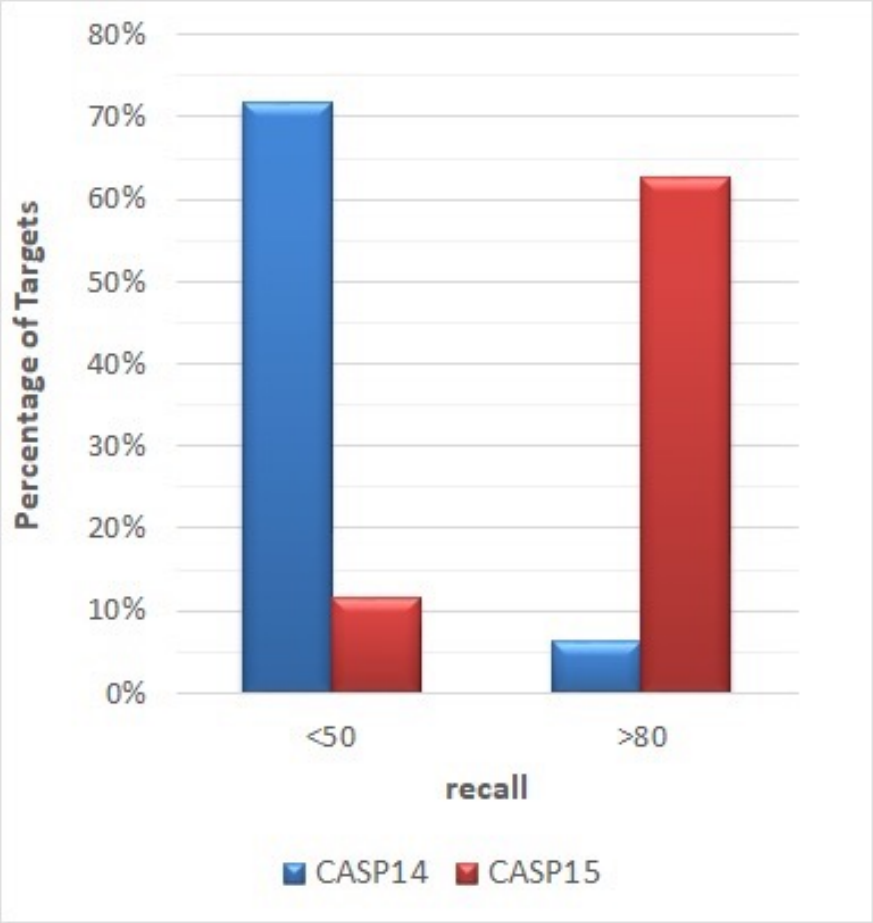


Figure 5: *% of CASP14 (blue) and CASP15 (red) protein assembly targets with high quality (>0.8) and low quality (<0.5) computed models as measured by the average contact precision (left) and recall (right)* (11). *In CASP15 the fraction of high-quality models increased from less than 10% to more than 60% by both measures. Correspondingly, the fraction of poor-quality models dropped precipitously.*

As figure 5 shows, the improvement in performance compared to the previous CASP is dramatic, and attributable to deep learning methods displacing earlier classical docking and homology methods.

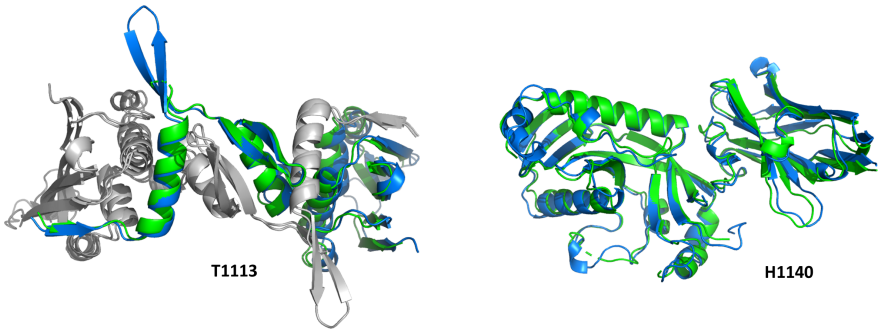


Figure 6 *Superposition of the experimental (green) and closest calculated structures (blue) for two protein complexes. T1113 is a phage shell homodimer with intertwined polypeptide chains at the interface. The left-*

hand subunit is shown in grey, to clarify the intertwined region. The interface score (ICS) is 93%. H1140 is a nanobody (right)-antigen (left) complex (ICS 81%).

Figure 6 shows two examples of high agreement with the corresponding experimental structures, each representing an assembly challenge class that was problematic for older methods: T1113 is a small bacterial homodimer with no homologous structures available. The two polypeptide chains intertwine across the interface in a domain-swap like manner, a feature that defeats classical docking methods. Deep learning methods treat the whole complex as a single entity, circumventing that difficulty. H1140 is a nanobody /protein antigen complex. Earlier benchmarking (29) had shown that immune complexes defeat at least the standard AlphaFold2-Multimer procedure. In CASP15 there are a total of eight immune complex targets (five nanobody complexes and three antibody complexes). Of these, two had homologous experimental complexes available, and so were easy targets. Three others have the lowest accuracy interfaces in this CASP (ICS 0.12, 0.30, and 0.45). But for the remaining three, high quality (ICS 0.74, 0.80, 0.81) models were produced by a small number of participating groups. Standard AlphaFold Multimer with default parameters was not effective on any of these, in accordance with the general observation that for many targets, enhanced sampling is necessary to obtain the best results.

As with the single protein structure category, the most effective methods in assembly modeling are based on AlphaFold2, usually the newer AlphaFold-multimer (31), a version of AlphaFold where training included data for protein complexes. Three of the methods are also in the top five performers in the single protein category. Again, as with the single protein category, the most successful groups used modifications of standard AlphaFold procedures, including much more extensive sampling through variations on MSA construction, the use of multiple seeds, an increased number of recycles and extensive network dropout. In addition, one group (32) devised a machine learning/Voronoi polyhedral interface scoring function which evidently aided in selection of accurate models. Details of methods can be found in the CASP15 assemblies assessment paper (11) and papers by some of the best performing groups.

Although the improvement in accuracy is enormous, there are still a substantial fraction of poor scoring interfaces. There are multiple possible reasons for the lag in performance. This is the first-time deep learning methods have been extensively used for protein complexes whereas this was the third CASP where deep learning has been used for single proteins. Thus, we may see substantial improvement next time as lessons are learned from CASP15. More fundamentally, there are many fewer structures of complexes in the PDB than single proteins, so that training set is inherently smaller. It may be possible to use the current methods to generate additional synthetic training data (33). Analogously to single proteins, interface accuracy probably falls off with the depth of the multiple sequence alignment spanning the interface (although one leading group reported omitting these data (34)). That may explain the generally weak performance for immune complexes, so it is encouraging to see partial success there.

2.4 | Estimating accuracy

A key attribute of experimental methods for determining macromolecular structure is that they return a generally reliable estimate of accuracy at the individual amino acid sequence level. In order to be taken seriously, calculated structures must also provide this information. CASP has long assessed self-accuracy estimates and also estimates by third parties who have developed methods for this purpose.

In CASP14 (2020), the new deep learning methods, especially AlphaFold2, provided models with very reliable per residue accuracy estimates, expressed as plDDT, the predicted lDDT. (lDDT is a metric reflecting the accuracy of a residue's environment in terms of the difference between experimental and calculated interatom distances (35)). In CASP14 it also became clear that third-party accuracy estimates are now generally less reliable than self-estimates and vary in reliability depending on the method used to build a model. So, in this CASP, assessment of self-accuracy estimates for tertiary structures is included, but third-party methods have been dropped. The assessor for this category showed that for single structures self-estimates of accuracy continue to be overall highly reliable, although more analysis is needed to determine how the reliability varies with circumstances (12).

For multimers, third-party selection of models appears impressive, with about 2/3 of targets having a loss of accuracy less than 0.1 in TM score (36) when models estimated to be most accurate are selected. Over half have a loss of less than 0.05. However, some of the methods use consensus over many models to estimate accuracy, rarely possible in practice. Amongst the best is a 'naïve' control method, suggesting the sophistication of other consensus methods is not adding much overall. But some publicly available methods requiring at most a few additional models rank highly and may be valuable for users, for example (32), (37) and (38) in this issue. Less satisfying is that, judging by the Pearson correlation between estimated and actual accuracy, ranking of accuracy across the full ranges of models is sometimes poor.

Self-accuracy estimates for multimers are only available in the form of submitted per residue plDDT values. The assessor provides average differences of these from actual lDDT values for core, surface, and interface residues. Overall, average differences are small for the best methods: less than 0.1 for core residues, but somewhat higher (0.16) for interface ones.

2.5 | Macromolecular Ensembles

Previously, CASP has assumed that there is only one relevant structure of a protein to compute, and that structure is represented by a single experimental structure. Increasingly, experimental methods provide more than one structure and computational methods should be able to produce all of these. In this CASP, we introduced an ensembles category with targets that have multiple experimentally observed conformations. As the ensembles evaluation paper discusses (9), there are still very limited experimental data of this type and even when there are data it is not always clear what a computational method should reasonably be expected to deliver. Nevertheless, the CASP15 results do provide preliminary insight into the abilities of computational methods in this area.

There are a number of encouraging results. Of particular note are reproduction of a domain-swap conformational change induced by a single mutation, sampling of three different conformations of an ABC transporter in the presence of different ligands; sampling of kinase substructures observed under different conditions; and sampling of alternative conformations of a small protein found in two different crystal forms. On the other hand, modeling of an RNA folding intermediate proved too difficult, as did different states of a Holliday junction complex. However, these failures may reflect the intrinsic difficulties of the targets as much as sampling deficiencies. The successful methods again used variants of AlphaFold2, with enhanced sampling, including adjustment of the multiple sequence alignment. In this respect, strategies are similar to those used for tertiary structure and protein assembly modeling, though details differ. Many preprints are appearing on improved ensemble modeling methods, so that we expect substantial advances in the next CASP round.

2.6 | Protein-ligand complexes

Accurate computation of binding modes for organic ligands is of great practical relevance because of its role in rational drug design. There have been several previous challenge experiments in this area, most notably the D3R series (39), but all have ceased operation. Ideal targets are likely those related to drug development, but in this first CASP round, only a small number of endogenous ligands were available. Additionally, participants were not provided with structures for the host protein (and one RNA) structures so that if a successful model of the host macromolecule was not generated, ligand docking was doomed to failure. Thus, this should be regarded as a pilot experiment and not a definitive assessment of the state of the art. Nevertheless, some useful insights were obtained by the assessor (14). First, participating deep learning methods were not as effective as traditional ones. Second, of the traditional methods, those that worked from homology or analogy rather than de novo docking were most successful. These are methods in which the PDB is searched for similar ligands and binding pockets, providing a starting point for pose refinement. Perhaps not surprisingly, less flexible ligands tended to more tractable.

2.7 | RNA structure

For a number of years, RNA Puzzles (6) has evaluated RNA modeling methods on a rolling basis as experimental structures became available. The rate of RNA experimental structure determination has increased

markedly recently, partly because of advances in cryo-electron microscopy. Consequently, it is now possible to obtain enough targets to make a CASP category for this area possible. The CASP15 RNA category was conducted in close collaboration with RNA Puzzles, and assessed by joint team led by Rhiju Das (for CASP) and Eric Westhof (for RNA Puzzles) (13). There were 12 targets in all, ranging in size from 30 bases to 720 bases, and including four designed structures. Two of the structures are of the same RNA molecule with a different number of partner proteins bound.

The assessors judged that except for the two RNA-protein complexes, the overall topology was correct for at least some models of each target, and Watson-Crick base pairing was also largely accurate, as was the relative positioning of the double helical regions. However, non-canonical pairing and irregular conformations regions were less successfully modeled. They also judged that most targets display evidence of significant flexibility, complicating assessment. For one target, the experimental group generated four different cryo-EM maps. The highest agreement of any submitted model was to one of the secondary density maps (Figure 7), illustrating that for some RNA structures additional experimental work may be needed for a proper assessment of computational models.

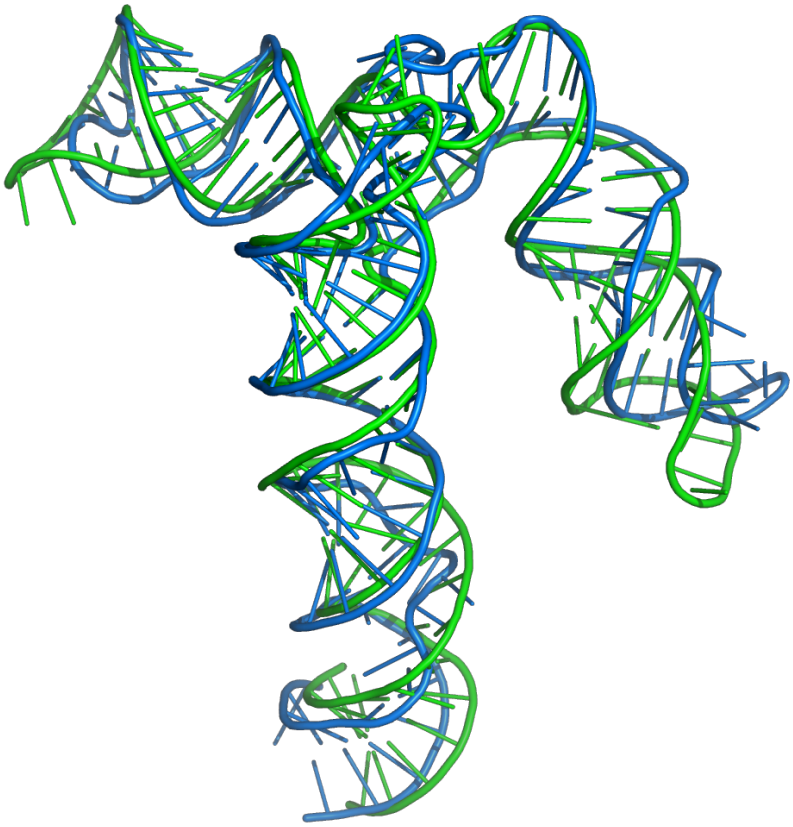


Figure 7: *Superposition of the closest model to experiment (blue) onto a structure of target R1156 (green), a homolog of the SARS-CoV-2 SL5 domain. This target showed substantial experimental flexibility, and the superposition is to the second highest resolution cryo-EM map. The GDT_TS is 51%.*

The most successful methods used classical RNA modeling methods, and although deep learning methods did take part, these were not yet competitive with the more established approaches. No good models were

submitted for the two protein-RNA complexes, and it appears these were beyond the scope of then available methods. However, newer methods have appeared since then that may be more capable, for instance (25).

3 | DISCUSSION

This CASP saw consolidation of the major progress in computing the structure of single proteins achieved in the previous round and extension of deep learning methods to other challenges in structural biology with impressive success for protein assemblies.

For single protein structures, the most effective methods obtained only slightly higher levels of accuracy to those of CASP14. It's not surprising that there is no major advance here, since for many targets agreement with experiment was likely already within experimental uncertainty, leaving little room for improvement. But there are some limitations to the methods. Shallow sequence alignments sometimes result in poor quality structures. Large proteins (more than about 1000 amino acids) may also have models of domains and domain interfaces that are of slightly lower accuracy than usually achieved.

All the most successful single protein methods used variations on AlphaFold2, sometimes embedding all or part of that software into their existing pipelines. The next most successful method is RosettaFold (24), though the difference in performance is substantial. RosettaFold2 is now available (25) and benchmarking shows improved performance. Several Large Language Models (LLMs) were included in this CASP, but performance lagged very significantly, including for shallow alignment targets where these methods were expected to be superior. Poor performance of these methods in this CASP round should not be seen as a final result, and the LLM methods may evolve to be more powerful.

Notably, running the standard version of AlphaFold2 with default parameters either through the ColabFold server (22) or locally installed often resulted in less accurate models. The primary reason for that appears to be that for some targets more extensive sampling than the default is needed. Successful groups used different approaches to increase sampling as well as tuning of multiple sequence alignments. As yet there are no benchmark studies that help a user chose between these approaches so as to optimize accuracy for the least increase in computing costs. Users are advised to carefully read the protocol descriptions in the relevant papers in this Proteins special issue and to keep an eye on the literature for useful studies. The key message is not to expect the best results using the standard protocol: if the structure obtained that way has acceptable predicted accuracy, usually no additional steps are necessary. However, if it falls short, and additional computing resources are available, further sampling is recommended.

The major advance this CASP was in the application of deep learning methods to protein assemblies. For many targets, but not all, agreement with experiment may be approaching experimental uncertainty limits, but it should be emphasized that we do not yet have adequate calibration of what that is. There were few transient complexes included in the target set and most of those were for antibody or nanobody-antigen complexes. Three non-homology complexes of this type were accurately modeled, three were not, an encouraging result compared to earlier benchmarking (29) but the methods do still have a way to go. Other apparent limitations are for some interfaces in large targets, although the role of experimental uncertainty in those results is unclear.

As with single proteins, the successful methods for protein complexes had AlphaFold2 at their core but used extensive sampling beyond the defaults. Several of the most successful groups were also successful for single targets, and the methods overlap. Also similarly to single protein targets, further benchmarking is needed to guide user protocol selection.

CASP has placed a long-running emphasis on not only the production of accurate protein structures but also on the provision of reliable estimates of co-ordinate uncertainty, with estimates provided both by participants submitting the structures (self-estimates) and by other groups who specialize in this area (3^{rd} party estimates). The increase in accuracy of protein complexes in this CASP brought new importance to accuracy estimates for these structures too. AlphaFold2-based tertiary structure self-accuracy estimates have high reliability, consistent with the 2020 results (40). For protein assemblies, third party accuracy estimates

will usually allow selection of a high-quality model, though not necessarily the best. Residue accuracy self-estimates correlate strongly with experiment, though are somewhat less reliable for interface residues than elsewhere.

CASP introduced three new categories this round, based on increased interest in the potential for deep learning methods to lead to advances there (17). The first is for modeling multiple structures in ensembles, both protein and RNA. Difficulties in obtaining suitable targets limited the significance of the results, but nevertheless, those that were obtained are generally encouraging. As with single proteins and protein complexes, successful methods were AF2-based, and incorporate a variety of enhanced sampling techniques. Although far from perfect, the results show that ensemble modeling is possible with current deep learning approaches. From the CASP perspective, the main difficulty going forwards is in obtaining suitable ensemble targets. We urge those who will have suitable experimental structures to get in touch.

Assessment of protein-ligand complexes is another new category in this CASP. Ideal targets here are series of compounds binding to the same protein. In spite of target limitations, a clear finding is that classical ligand docking methods were still superior to the new deep learning methods. These results are now a year old, and improved methods are no doubt under development.

The final new modeling category is RNA structure (13). Here, in spite of promising new deep learning methods having been published (7) and a number of CASP15 groups exploring the approach, classical approaches proved superior. Overall fold accuracy, Watson-Crick helical regions and their packing is often accurately modeled, but there were difficulties with modeling the more irregular regions of the targets. The effects of flexibility also complicate the assessment. The greater flexibility of RNA structures suggests that current experimental procedures may not always be adequate for assessment of computational methods. Conversely, computational methods must produce ensembles of structures to adequately represent experimental reality.

This round of CASP saw one very major advance (greatly improved accuracy in modeling protein complexes) and greater clarity on the application of AF2 to single proteins, as well as interesting and provocative starts to modeling of macromolecular complexes, protein ligand complexes, and RNA structure. We look forward to further major gains in the next CASP, in 2024.

ACKNOWLEDGEMENTS

CASP depends on major contributions from the experimental community in providing targets, the work of assessors and their teams, and the participants who submit computed structures. We are grateful to 48 experimental groups from 14 countries who provided targets for CASP15. We thank the assessment teams of Dan Rigden (tertiary structure), Ezgi Karaca (protein assemblies), Gabriel Studer (accuracy estimation), Rhiju Das (RNA structure) and Pat Walters (protein-ligand complexes). We are also grateful to Marc Lensink and Shoshana Wodak for the CAPRI assessment of protein assemblies and Eric Westhof for the RNA Puzzles assessment work. Within out the willingness of the 89 research groups who submitted computed structures, there could be no CASP. We thank Arne Elofsson and Claudio Mirabello for providing standard AlphaFold2 results. The Center for CASP at UC Davis is supported by a grant from the US National Institute of General Medical Sciences (NIGMS/ NIH), R01GM100482 to KF.

CONFLICT OF INTEREST

The authors declare they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data used in this work are freely available on the CASP web site.

ORCID

Andriy Kryshtafovych https://orcid.org/0000-0001-5066-7178

Torsten Schwede https://orcid.org/0000-0003-2715-335X

Maya Topf https://orcid.org/0000-0002-8185-1215 Krzysztof Fidelis https://orcid.org/0000-0002-8061-412X

John Moult https://orcid.org/0000-0002-3012-2282

REFERENCES

1. Anfinsen CB. Principles that govern the folding of protein chains. Science. 1973;181(4096):223-30.

2. Robin X, Haas J, Gumienny R, Smolinski A, Tauriello G, Schwede T. Continuous Automated Model EvaluatiOn (CAMEO)-Perspectives on the future of fully automated evaluation of structure prediction methods. Proteins. 2021;89(12):1977-86.

3. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-Round XIII. Proteins. 2019;87(12):1011-20.

4. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-Round XIV. Proteins. 2021;89(12):1607-17.

5. Wodak SJ, Velankar S, Sternberg MJE. Modeling protein interactions and complexes in CAPRI: Seventh CAPRI evaluation meeting, April 3-5 EMBL-EBI, Hinxton, UK. Proteins. 2020;88(8):913-5.

6. Magnus M, Antczak M, Zok T, Wiedemann J, Lukasiak P, Cao Y, et al. RNA-Puzzles toolkit: a computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools. Nucleic Acids Res. 2020;48(2):576-88.

7. Townshend RJL, Eismann S, Watkins AM, Rangan R, Karelina M, Das R, et al. Geometric deep learning of RNA structure. Science. 2021;373(6558):1047-51.

8. Crampon K, Giorkallos A, Deldossi M, Baud S, Steffenel LA. Machine-learning methods for ligand-protein molecular docking. Drug Discov Today. 2022;27(1):151-64.

9. Kryshtafovych A, Montelione G, Rigden D, Mesdaghi S, Karaca E, Moult J. Breaking the conformational ensemble barrier: Ensemble structure modeling challenges in CASP15. Proteins. 2023 (this issue, https://doi.org/10.1002/prot.26584).

10. Simpkin AJ, Mesdaghi S, Sanchez Rodriguez F, Elliott L, Murphy DL, Kryshtafovych A, et al. Tertiary structure assessment at CASP15. Proteins. 2023 (this issue. https://doi.org/10.1002/prot.26593).

11. Ozden B, Kryshtafovych A, Karaca E. The Impact of AI-Based Modeling on the Accuracy of Protein Assembly Prediction: Insights from CASP15. Proteins. 2023 (this issue, in production).

12. Studer G, Tauriello G, Schwede T. Assessment of the assessment - All about complexes. Proteins. 2023 (this issue, in production).

13. Das R, Kretsch RC, Simpkin A, Mulvaney T, Pham P, Rangan R, et al. Assessment of three-dimensional RNA structure prediction in CASP15. Proteins 2023 (this issue, in production).

14. Walters P, Robin X, Studer G, Durairaj J, Eberhardt J, Schwede T. Assessment of Protein-Ligand Complexes in CASP15. Proteins. 2023 (this issue, in production).

15. Alexander LT, Durairaj J, Kryshtafovych A, Abriata LA, Bayo Y, Bhabha G, et al. Protein target highlights in CASP15: Analysis of models by structure providers. Proteins. 2023 (this issue, http://doi.org/10.1002/prot.26545).

16. Kretsch RC, Andersen ES, Bujnicki JM, Chiu W, Das R et al. RNA target highlights in CASP15: Evaluation of predicted models by structure providers. Proteins. 2023 (this issue, https://doi.org/10.1002/prot.26550).

17. Kryshtafovych A, Antczak M, Szachniuk M, Zok T, Kretsch RC, Rangan R, et al. New prediction categories in CASP15. Proteins. 2023 (this issue, http://doi.org/10.1002/prot.26515).

18. Kryshtafovych A, Rigden DJ. To split or not to split: CASP15 targets and their processing into tertiary structure evaluation units. Proteins. 2023 (this issue, http://doi.org/10.1002/prot.26533).

19. Zemla A, Venclovas, Moult J, Fidelis K. Processing and evaluation of predictions in CASP4. Proteins. 2001;Suppl 5:13-21.

20. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2023.

21. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res. 2003;31(13):3370-4.

22. Mirdita M, Schutze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022;19(6):679-82.

23. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Applying and improving AlphaFold at CASP14. Proteins. 2021 Dec;89(12):1711-1721.

24. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373(6557):871-6.

25. Baek M, Anishchenko I, Humphreys IR, Cong Q, Baker D, DiMaio F. Efficient and accurate prediction of protein structure using RoseTTAFold2. bioRxiv. 2023:2023.05.24.542179.

26. Korgaonkar A, Han C, Lemire AL, Siwanowicz I, Bennouna D, Kopec RE, et al. A novel family of secreted insect proteins linked to plant gall development. Curr Biol. 2021;31(9):2038.

27. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9.

28. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv. 2022:2022.07.20.500902.

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

30. Lensink MF, Brysbaert G, Raouraoua N, Bates P, al. E. Impact of AlphaFold on Structure Prediction of Protein Complexes: The CASP15-CAPRI Experiment. Proteins. 2023 (this issue, in press).

31. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer. bioRxiv. 2022:2021.10.04.463034.

32. Olechnovic K, Valancauskas L, Dapkunas J, Venclovas C. Prediction of protein assemblies by structure sampling followed by interface-focused scoring. Proteins. 2023 (this issue, https://doi.org/10.1002/prot.26569).

33. Savage N. Synthetic data could be better than real data. Nature. 2023, Apr 27 (doi: 10.1038/d41586-023-01445-8).

34. Peng Z, Wang W, Wei H, Li X, Yang J. Improved protein structure prediction with trRosettaX2, AlphaFold2, and optimized MSAs in CASP15. Proteins. 2023 (this issue, https://doi.org/10.1002/prot.26570).

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

36. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins. 2004;57(4):702-10.

37. Edmunds NS, Alharbi SMA, Genc AG, Adiyaman R, McGuffin LJ. Estimation of model accuracy in CASP15 using the ModFOLDdock server. Proteins. 2023 (this issue, https://doi.org/10.1002/prot.26532).

38. Liu J, Liu D, He G, Zhang G. Estimating protein complex model accuracy based on ultrafast shape recognition and deep learning in CASP15. Proteins. 2023 (this issue, https://doi.org/10.1002/prot.26564).

39. Parks CD, Gaieb Z, Chiu M, Yang H, Shao C, Walters WP, et al. D3R grand challenge 4: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies. J Comput Aided Mol Des. 2020;34(2):99-119.

40. Kwon S, Won J, Kryshtafovych A, Seok C. Assessment of protein model structure accuracy estimation in CASP14: Old and new challenges. Proteins. 2021;89(12):1940-1948.