Tertiary structure assessment at CASP15

Adam J. Simpkin[1], Shahram Mesdaghi[1,2], Filomeno Sánchez Rodríguez[1,3,4], Luc Elliott[1], David L. Murphy[1], Andriy Kryshtafovych[5], Ronan M. Keegan[6], Daniel J. Rigden[1] *

[1] Institute of Structural, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 7ZB, England

[2] Computational Biology Facility, MerseyBio, University of Liverpool, Crown Street, Liverpool L69 7ZB, United Kingdom.

[3] Life Science, Diamond Light Source, Harwell Science and Innovation Campus, Didcot, Oxfordshire OX11 0DE, England

[4] York Structural Biology Laboratory, Department of Chemistry, University of York, Heslington, York, England

[5] Genome Center, University of California, Davis, California

[6] UKRI-STFC, Rutherford Appleton Laboratory, Research Complex at Harwell, Didcot OX11 0FA, England

*Correspondence e-mail: drigden@liverpool.ac.uk

Abstract

The results of tertiary structure assessment at CASP15 are reported. For the first time, recognising the outstanding performance of AlphaFold 2 (AF2) at CASP14, all single chain predictions were assessed together, irrespective of whether a template was available. At CASP15 there was no single stand-out group, with most of the best-scoring groups - led by PEZYFoldings, UM-TBM and Yang Server - employing AF2 in one way or another. Many top groups paid special attention to generating deep Multiple Sequence Alignments (MSAs) and testing variant MSAs, thereby allowing them to successfully address some of the hardest targets. Such difficult targets, as well as lacking templates, were typically proteins with few homologues: small size, high α-helical content and monomeric structure were other possible aggravating factors. Local divergence between prediction and target correlated with localisation at crystal lattice or chain interfaces, and with regions exhibiting high B-factor factors in crystal structure targets, and should not necessarily be considered as representing error in the prediction. However, analysis of exposed and buried side chain accuracy showed room for improvement even in the latter. Nevertheless, a majority of groups produced high quality predictions for most targets which are valuable for experimental structure determination, functional analysis and many other tasks across biology. These include those applying methods similar to those used to generate major resources such as the AlphaFold Protein Structure Database and the ESM Metagenomic atlas, where confidence estimates were also notably accurate.

1 | Introduction

For nearly 30 years the Critical Assessment of Structure Prediction (CASP) experiments have monitored, assessed and incentivised developments in protein structure prediction [1,2]. Every two years, predicting groups are invited to model protein sequences in advance of their experimental structures - be they determined by X-ray crystallography, cryo-Electron Microscopy or Nuclear Magnetic Resonance - becoming publicly available. Independent groups then assess performance using standardised metrics and statistical models in order to rank groups by performance at each event. More importantly, the exercise also serves to publicise progress and stand-out methods to the broader communities who benefit from protein modelling, and to point to areas requiring further improvement.

Central to the CASP endeavour since the beginning has been assessment of structural modelling of single protein chains. Even in the earliest CASP exercises, simpler targets - those for which a homologous structure could be identified in the Protein Data Bank [3] - were modelled quite well. In contrast, *ab initio* modelling (aka *de novo* or template-independent modelling) has seen dramatic progress from very poor performance in the early days [4], via increasingly sophisticated fragment assembly methods that made good models of small proteins [5] to the modern era of Machine Learning and especially Deep Learning [6]. At CASP14, AlphaFold 2 (AF2) emerged as the top-performing method, by some distance, in both the template-based [7] and *ab initio* [8] categories. Importantly, the performance on hard targets was close to that seen on easier targets [2] rendering unnecessary the twin-track assessment. Hence, at CASP15, the organisers united the previous two categories into one assessment of modelling of single protein chains.

The results of the CASP15 assessment of single chain modelling are presented here. There were submissions from 132 groups for 112 evaluation units deriving from 77 single chain targets. We evaluated the performance of the 118 groups which had submitted models for at least 10 of the 112 evaluation units. In contrast to CASP14 there was no single stand-out group: many groups performed very well on a majority of targets with the leading groups distinguished by their ability to tackle the hardest targets. Most of the best-performing groups used AF2 in one way or another. The two submissions using protein Language Model (pLMs) suggest that such methods are not yet competitive with those using Multiple Sequence Alignments (MSAs), and claimed benefits for targets with no or few homologues [9,10] were not apparent on the CASP15 targets. Nevertheless, many models, including those derived by pLM-based methods can, with appropriate and sometimes essential editing, solve most crystal structures.

2 | Materials and Methods

2.1 Target definition

The procedures for processing full-length CASP15 targets into evaluation units (EUs) are described in detail elsewhere [11]. Briefly, protein chains were split into compact structural units (here referred to simply as domains) using the results of several different analytic methods, and also considering similarities to other protein structures. These domains were combined into EUs where a majority of groups successfully captured their relative orientation. Where, on the other hand, groups predicted the individual structural units well but a majority failed to predict their packing, the individual structural domains were retained as EUs. Overall, 112 EUs were derived from 77 tertiary structure prediction targets. Three EUs - T1114s1-D2, T1157s1-D2 and T1157s1-D3 - were not evaluated because of the low resolution of the cryo-EM maps in their local areas.

As in previous CASPs, EUs were assigned to target difficulty categories: TBM (template-based modelling, easy or hard), FM (free modelling), and the TBM/FM overlap category. Unlike previous CASPs, this procedure was done automatically using methods designed to recapitulate, as far as possible, previous assignments that had significant manual input [11]. Ultimately, there were 47 EUs in the TBM-easy class, 15 TBM-hard, 8 TBM/FM, and 39 in the FM category. This last number is significantly larger than seen in recent CASPs.

2.2 Scoring and ranking

Following the practice established by previous CASPs, the group ranking was done using a composite score including metrics relating to global fold correctness, main chain quality, side chain accuracy and the accuracy of the confidence estimates. Z-scores were employed to make all measures dimensionless and to represent relative, rather than absolute, performance across all measures in a uniform way. The CASP15 score was modified from the CASP14 predecessor to include reLLG values. The CASP14 score [7] was:

$$S_{CASP14} = \left(\frac{1}{16}\left(Z_{LDDT} + Z_{CADaa} + Z_{SG} + Z_{sidechain}\right) + \frac{1}{12}\left(Z_{MolPrb-clash} + Z_{backbone} + Z_{DippDiff}\right) + \frac{1}{4}\left(Z_{GDT\_HA} + Z_{ASE}\right)\right)$$

Here, the set of metrics in the first parenthesis focuses on local and side chain quality: *LDDT* is the local Difference Distance Test that evaluates the agreement between the all-atom distance maps of target and model[12] , *CADaa* is the all atom variant of the CAD score looking at residue contact surface areas[13] , *Sphere-Grinder (SG)*, measures how well the model captures the local atomic environments of each residue[14] , and *sidechain* refers to one of the two torsion angle deviation metrics introduced for template-based model assessment at CASP13[15]. The second group focuses more on main chain quality. It includes *MolPrb-clash*, which refers to the number of serious atom clashes detected by MolProbity [16] (the calculation also involves side chains), *backbone*, the second, main chain-focused of Croll et al's torsion angle deviation metrics [15] , and *DipDiff*, which measures interatomic distances involving Cα and O atoms between neighbouring residues and compares them between target and model [17]. In the third group are *GDT_HA*, the high-accuracy variant of the Global Distance Test - Total Score (GDT_TS) which measures global fold accuracy [18] and *ASE*, the Accuracy Self Estimate, measuring the correlation between error estimates and actual model errors.

For CASP15, we implemented two modifications to the ranking score. First, the ASE measure was calculated on atomic predicted LDDT values (pLDDT) instead of predicted coordinate errors in Ångstroms [19], and second, the reLLG measure [20] was added to the scoring. The change in the ASE calculation was dictated by the change in the prediction requirements for CASP15, where predictors were asked to estimate accuracy of atoms' placements in a model in terms of pLDDT and not the distance as in CASP14. The inclusion of the reLLG is rationalised by its modest correlation with other elements of the ranking score (the highest pairwise correlation coefficient was 0.69 with GDT_HA) and the importance of protein modelling for its widely-used downstream application in the experimental protein crystallography. Conceptually, the reLLG is a coordinates-only score predicting the usefulness of a model for Molecular Replacement (MR)[20]. We included it in the ranking formula with the same weight as GDT_HA and ASE.

The CASP15 score was:

$$S_{CASP15} = \left(\frac{1}{16} \left(Z_{LDDT} + Z_{CADaa} + Z_{SG} + Z_{sidechain} \right) + \frac{1}{12} \left(Z_{MolPrb-clash} + Z_{backbone} + Z_{DippDiff} \right) + \frac{1}{6} \left(Z_{GDT\_HA} + Z_{ASE} + Z_{reLLG} \right) \right)$$

We considered changing the assessment of side chain accuracy. However, additional potential metrics were found to correlate too strongly to the existing side chain torsion angle deviation metric to justify inclusion. For example, the GDC_SC measure, a Global Distance Calculation for side chains, had a correlation coefficient of 0.95 with the GDT_HA metric that was already part of the composite score. Similarly, the Average Absolute Accuracy (AAA) measurement from the SCWRL4 package[21] had a correlation coefficient of -0.96 with the existing side chain torsion angle score.

As mentioned, z-scores were used instead of raw scores and, as is also customary at CASP, the calculations proceeded in two rounds. In the first, models scoring below an initial z-score of -2 were considered as outliers and excluded. Z-scores were then recalculated, but only positive z-scores included in the ranking calculations. In this way groups who, perhaps through more speculative and experimental methods, produced a few dramatically poor models were not prevented from having consistently good performance recognised elsewhere. Again as previously, the final rankings were based on sums of z-scores in order to reward groups performing well across all targets. Finally, it is to be noted that the composite scoring addresses previous concerns that models may capture an overall global fold correctly, but perform poorly in other regards [15]. Nevertheless, since the accuracy of the fold is the primary consideration, relative performance of groups on GDT_HA alone was also examined.

Calculations were done using a modified version of the code repository created for CASP14 by Joana Pereira and are available at https://github.com/hlasimpk/CASP15_high_accuracy.

Multiple Sequence Alignment (MSA) depth was measured as Neff/length with data kindly provided by Claudio Mirabello from the National Bioinformatics Infrastructure Sweden at SciLifeLab (doi.org/10.17044/scilifelab.22769996). Calculations were based on alignments that were generated by the public AlphaFold2 server with default parameters and databases (https://github.com/clami66/AF_server/). Classification of targets using DSSP [22,23] defined them as all-α (100% of regular secondary structure was α-helix), mostly α (>65% α), mixed, mostly β (>65%) or all-β (100% of regular secondary structure was β-sheet).

2.3 Factors affecting local model quality

In order to determine whether local modelling errors were more likely to be found in the vicinity of intermolecular interfaces, predicted models were analysed using the procedure described in the CASP14 refinement assessment study [24]. Only higher quality (GDT_TS >80) model_1 submissions were included in the analysis so that errors could be considered local, rather than as resulting from overall global poor performance. Error was assessed and compared to structural context at both residue and 'region' levels

- For residue level analysis, the LGA distance (between the predicted model and experimental structure superimposed using the sequence-dependent algorithm) was compared for residues contributing to crystal, chain or domain interfaces or to none of these. Residues were considered at a crystal, chain or domain interface where they had at least three < 10Å Cα-Cα contacts with residues in neighbouring symmetry mates, chains or domains, respectively.

- Error regions were defined as follows. A five residue-window rolling average LGA distance (defined as above) was calculated for each residue in the predicted models. Error regions were then defined as comprising at least three consecutive residues with a rolling LGA average of at least 3Å. These error regions were then defined as being at a crystal lattice interface, a chain interface or a domain interface if the residues within the region had an average of at least 0.5 residues within a radius of 10Å in a symmetry mate, another chain or a different domain, respectively [24]. Again, distances were measured between Cα atoms.

To assess the relationship between B-factors and local error, for selected groups, residues in higher-quality (GDT_TS >80) model_1 submissions were analysed. Within each target, residue B-factors were first normalised. All residues from all targets were then combined and residues placed into 10 equally sized bins according to normalised B-factors.

2.4 Side chain assessment

SCRWL4's AAA sidechain score [21] measures the percentage of the model's χ-angles for each residue that are within 40° of their corresponding angles in the reference structure. A score was calculated for each residue and then averaged over surface and non-surface side chains for the top-scoring model (by GDT_HA) for each target. To define surface and non-surface residues, the Shrake-Rupley algorithm [25] was used. Prior to the definition of EUs, the solvent accessibility of each target residue was calculated, and a residue was considered part of the surface if its solvent accessibility was greater than 20 percent [26].

In addition, the torsion angle deviation metrics for side chain and main chain [15] were plotted against each other as contour maps in order to assess the dependence of side chain placement accuracy on main chain modelling.

2.5 Scoring against X-ray crystallographic data and Molecular Replacement

2.5.1 Assessing the models' potential for success in Molecular Replacement

Models were tested against the experimental diffraction data, where available, by calculating LLGs for ideally placed structures and then, for selected groups, by attempting Molecular Replacement (MR). Details for the set of 17 targets with diffraction data are shown in the Supp Table 1. The top submitted model by each group, model_1, was first processed to remove residues with low pLDDT values. The current version of Slice'N'Dice [27] (SnD) removes residues for which the B-factor column, here recording pLDDT values, contains a value below 70 for the first atom encountered. Here that means that a few residues may have been discarded where their first atomic pLDDT was less than 70 but the mean across the residue exceeded that value (or kept where the reverse was true). However, since only 4 groups - namely Bench, Yang, Yang-Server and Yang-Multimer - chose to submit atomic, rather than per-residue, pLDDTs the impact of this was small. Note that no attempt was made here to correct entries from the handful of groups who apparently did not have pLDDT on the expected 0-100 scale in the B-factor column. The model was then placed ideally onto the target crystal structure using Gesamt [28] to do the structural superposition. Models were placed for all copies of the molecule in the asymmetric unit. The positioning of the model(s) was further optimised by using Phaser [29] to perform rigid refinement. This step also generated a total Log Likelihood Gain (LLG) for the placed model(s). A simple ranking of groups was generated by awarding the group responsible for the best model of a target x points, where x is the total number of groups having attempted a prediction for 1 or more of the 17 targets. The group producing the model with the second-best LLG was awarded x-1 points and so on. An LLG of 60 or more for the placement of the first component in an MR search is considered to be indicative of correct placement [30]. Models scoring less than this threshold did not receive any points. Groups that did not attempt a prediction for the target received no points. The ranking was on the total number of points across the 17 targets for which diffraction data were available.

The impact of domain splitting on alignment between models and target and resulting LLG values was also explored. Using the slice function of the SnD pipeline in CCP4, models were subjected to splitting into 2-4 rigid regions that might separately fit better to corresponding regions in the target structure. In order to process models of all origins uniformly, CCTBX's [31] PAE-based domain decomposition was not used: instead the purely coordinate-based Birch algorithm from SciKit-learn [32] was applied. These domain regions were then placed in the same way as described above, producing LLG values for all components matching the contents of the asymmetric unit.

2.5.2. Full Molecular Replacement

Model 1 from a subset of the top scoring groups from the alignment tests were used in a full MR test for each of the 17 targets. The same model from the ESM-single-sequence group was also used in these tests to assess how well models produced using the pLM-based method would perform in full MR. For this test, the unsplit models were used and subjected to B-factor conversion and with residues having a pLDDT below 70 removed. Phaser was used to carry out the MR with success measured by the resulting LLG.

The case of T1145 (636 residues) was explored in more detail since the presence of several domains made it likely that predicted models would differ from the conformational state captured in the crystal structure. Determining the structure using a predicted model is further complicated by the presence of two copies in the asymmetric unit. The resolution of the diffraction data is 2.2 Angstroms. SnD was used to attempt a structure solution using search models generated by splitting the model into 2, 3 and 4 pieces. MR was also attempted using the unsplit model. Results were refined with Refmac5 [33] before model rebuilding using Modelcraft [34]. Success was measured using the Rfactors achieved following the model rebuilding step.

2.6 Function prediction

Targets were selected based on their interpretability by structure-based methods. The selection was based on information given to the CASP predictors in combination with analysis and literature review of the targets. Four enzymes T1146 (a putative peptidoglycan hydrolase with a catalytic triad), T1110 (isocyanide hydratase with a catalytic dyad), T1127 (NATA1 with a catalytic dyad) and T1188 (chitinase with a catalytic triad) were selected to determine the accuracy of modelling of catalytic sites. The fit function of PyMOL was used to calculate all-atom rmsd values between catalytic sites in predictions and targets. The chemical equivalence of side chains for Asp, Gly and Tyr on 180° rotations about their Cβ-Cγ bonds was taken into account. Only models with GDT_HA >30 were considered in the analysis.

Three targets were identified as DNA binding proteins; T1153, T1170 and T1151. The three experimental structures were screened for predicted DNA-binding capability with both DNABIND [35] and BindUP [36]. Of the three targets only T1151 was predicted to be a DNA binding protein by either method so only models for T1151 were processed.

3 | Results

3.1 Overall group rankings

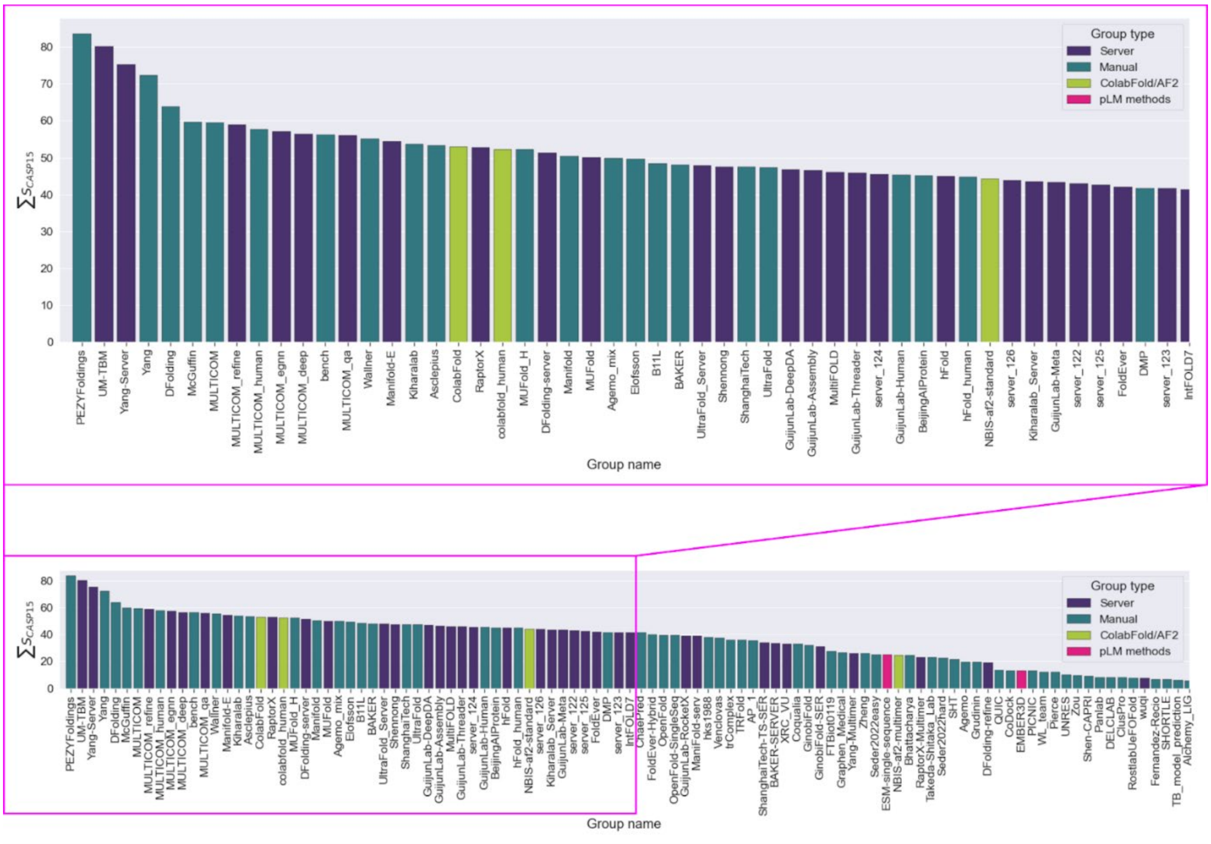


Figure 1. Cumulative group ranking on 109 CASP15 evaluation units. Groups are colour-coded indigo for server, i.e. a purely automated modelling protocol, and teal for manual where human intervention is allowed. Pure AlphaFold 2 comparison runs based on the original DeepMind protocol or its ColabFold version are shown in green. Pink is used for the two groups employing exclusively protein Language Model methods.

Fig 1 shows the CASP15 group ranking, expressed as the sum of the per-target S_{CASP15} scores, as defined in Methods. Comparison with the rankings according to the previous CASP formulae (Supp Fig 1) shows very similar results in the top positions. Several interesting conclusions can be drawn. Firstly, automated servers (indigo in Fig 1) are higher in rankings and appear better represented at the top of the performance table in CASP15 than in any previous CASP. Of the top three places in CASP15, two are occupied by servers, which never happened before in any of the prediction categories.

Secondly, it is clear that the 'control' runs of AF2, both in the original DeepMind implementation as of March 2022 (group names NBIS-AF2-standard and NBIS-AF2-multimer) [37] and a later ColabFold version (group names ColabFold and colabfold_human) [38], are a little way off the best performance possible, though each produces many high-quality models (see below). Detailed descriptions of the best groups' methods are to be found elsewhere in this issue, but common themes are paying special attention to the MSA depth by collecting sequences from different databases, and sampling across models resulting from MSAs that differ in their depth, database origin, or sub-clustering.

Thirdly, despite the moderate performance of AF2 controls, most of the best groups used AF2 in one way or another: the best performing method that was entirely independent of AF2 was that of the BAKER group, 28th place, using RoseTTAFold2 [39]. Among those methods using AF2, there was an interesting variety of strategies although many groups placed significant emphasis on generating models based on diverse MSAs and selecting the best models from the resulting sets. There was also an interesting trend towards creatively combining AF2 predictions with alternative predictive methodologies. Thus, the UM-TBM group [40] used AF2 predictions to guide replica exchange Monte Carlo simulations within the I-TASSER [41] framework, while the Yang-Server group [42] selected from AF2 and trRosettaX2 results for its submissions. Finally, the pLM-based methods (pink in Fig 1) are not currently competitive with MSA-based methods like AF2: the best-placed pLM group, ESM-single-sequence [43], is placed 74th among 118 evaluated participants.

Looking at the number of times each group produced the absolute best possible model (Supp Fig 2) yields a slightly different perspective. PEZYFoldings [44] remains in first place by both CASP15 and GDT scores, the top server method UM-TBM is in second place by all scores and Yang-Server places third by GDT_HA. However, DFolding (fifth in the overall ranking in Fig 1) rises to third by composite CASP15 and GDT_TS scores, while ShanghaiTech (31st in the overall ranking) rises to fourth place in the CASP15 table and fifth in the GDT_TS ranking.

3.2 Group performance on different targets

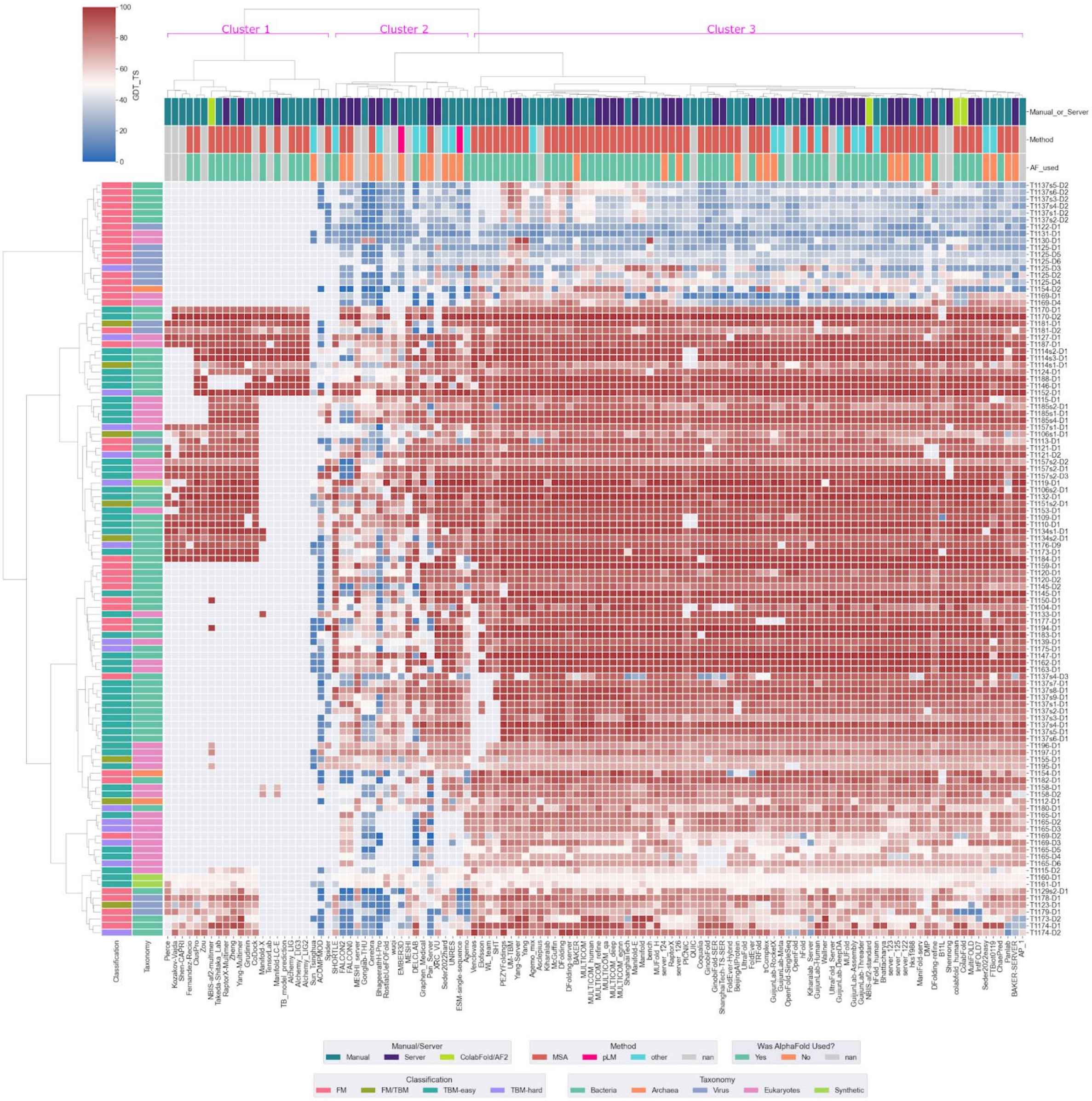


Figure 2. Ward's clustering applied to GDT_TS values (red good, blue bad, grey no submission) achieved by the 118 evaluated groups on the 109 Evaluation Units (EUs) that were considered. EUs are additionally annotated on the left with colour codings relating to classification (TBM_easy, TBM_hard, FM/TBM or FM) and taxonomy of the original target sequence (Bacteria, Archaea, Virus, Eukaryote, Synthetic). Groups are annotated on the top according to whether they were Server (indigo) or Human (blue), by broad category of method and according to if AF2 was used by a group. Note that the submitted Abstracts from some groups did not always allow confident inference of these aspects (grey labels). Three clusters of groups discussed in the text are indicated in magenta.

Fig 2 provides a two-dimensional heat map of CASP15 groups and targets clustered based on the GDT_TS score. Groups to the left in cluster 1 focussed solely on multimeric targets. Cluster 2 contains a number of groups whose mixed blue and red colouring indicates variable performance: notably, these methods tended to not employ AF2 (or it was unclear from the submitted abstract whether that was the case). Cluster 3 includes a broad swathe of groups, many based one way or another on AF2, that produced largely good to excellent models over most of the target EUs. Table 1, for example, shows that selected methods (comprising the best-performing MSA-based methods, AF2 controls and the best pLM-based method) produce model_1 predictions that have the correct topology (GDT_TS ≥ 45) in a large majority of cases. Remarkably, the best methods produce up to half of predictions with GDT_TS ≥ 90, a rough benchmark of differences between crystal forms of the same protein, and hence a reasonable ceiling on expected predictive performance. Again, two server groups that provide standardised versions of AF2: the NBIS-af2-standard, implementing a version of the original DeepMind AF2 protocol [37], and ColabFold [38] - perform somewhat worse than the top groups (Fig 2 and Table 1)

Looking next at the clustering by targets (vertical axis), there is a notable cluster of 18 at the top, for which average model quality was clearly lower, with most groups scoring GDT_TS <50. Strikingly, 17 out of these 18 EUs were FM targets (the exception being T1125-D3, classified as TBM-hard), illustrating how the traditional four-way classification of EUs by CASP [11], which is based on sequence and structural similarity of targets to PDB entries, still picks out targets which are more likely to prove difficult. However, it is important to note that another 22 FM targets were found in the main block of better-predicted targets. Thus, FM targets are clearly predisposed to difficulties but may, depending on other characteristics (see later), still be well-predicted. The other notable characteristic of the hard group of 18 was its richness in EUs derived from viral targets. Seven of the 18 are viral in origin, whereas the same is true only for six of the remaining 91 targets. This is a significant difference by Fisher's Exact test [45] ($P = 0.001$; $P < 0.05$). Inspection of Fig 2 confirms that the best-performing groups overall are those who also produced good models for the most challenging targets.

3.3 Target difficulty

3.3.1 What makes a target difficult

Factors possibly explaining the greater difficulty of some targets, even to the best-performing groups, were explored. The objective was to identify limitations of current methods and also to explain an apparent drop in overall performance by the best groups here compared to the best group at CASP14 (DeepMind with AlphaFold 2). The latter issue is also explored in the introductory paper of this journal issue [46].

Fig 3a illustrates a number of criteria that seem to link to target difficulty measured as median GDT_TS for the top 10 groups overall (Fig 3a). The most obvious characteristic of difficult targets is the availability of only a shallow multiple sequence alignment (MSA) when homologues are collected from sequence databases. MSA depth here is measured as Neff [47] normalised by length [48,49]. The hardest targets in Fig 3a all have very low Neff/length values: T1122, T1131 (the hardest target) and T1130 were all singletons in the principal public databases. The same trend persists in a graph calculated using models from all the groups (Supp Fig 3). Evolutionary covariance information extracted from the MSA is used by methods such as AF2 to predict contacts and distances between residues, and is particularly key for AF2 for the derivation of an initial structure model [37]. Where neither high-quality covariance information nor a structural template (i.e. for FM targets) is available then even the best groups may struggle. This feature can be related to the abundance of virus-derived targets in the most difficult group in Fig 2: as previously recognised, the rapid evolution of viral protein sequences can hamper the recognition and alignment of homologues in sequence databases [8].

Protein Language Models (pLMs) represent a new approach to structure prediction [9,10,43,50]. It has been asserted [9,10] that they may be less dependent on MSA-derived information and hence potentially capable of comparatively better performance than MSA-based methods on low Neff/length targets. Fig 3b, in which targets are ordered by their Neff/length, does not support this idea. The best pLM-based methods are competitive with, though generally not quite as good as, MSA-based methods for high Neff/length targets on the left. However, the performance deficit actually increases on the right as available MSAs become shallower.

Beyond low Neff, further analysis pinpoints other potential aggravating factors. For example, all of the five hardest targets are relatively small (Fig 3a), ranging in size from 66 to 234 residues, while no target larger than 300 residues achieved a median GDT_TS of less than 67. Similarly, the same five targets are either all-α or mainly α by secondary structure composition and there seems to be a tendency for the more α-rich targets to extend to lower median GDT_TS values among the results of the top 10 groups (Fig 3a). Finally, although numbers are small, it is interesting to note that four of the five hardest targets derive from crystal structures: the example of T1122 (below), though extreme, illuminates the particular complications that X-ray crystallography may sometimes introduce.

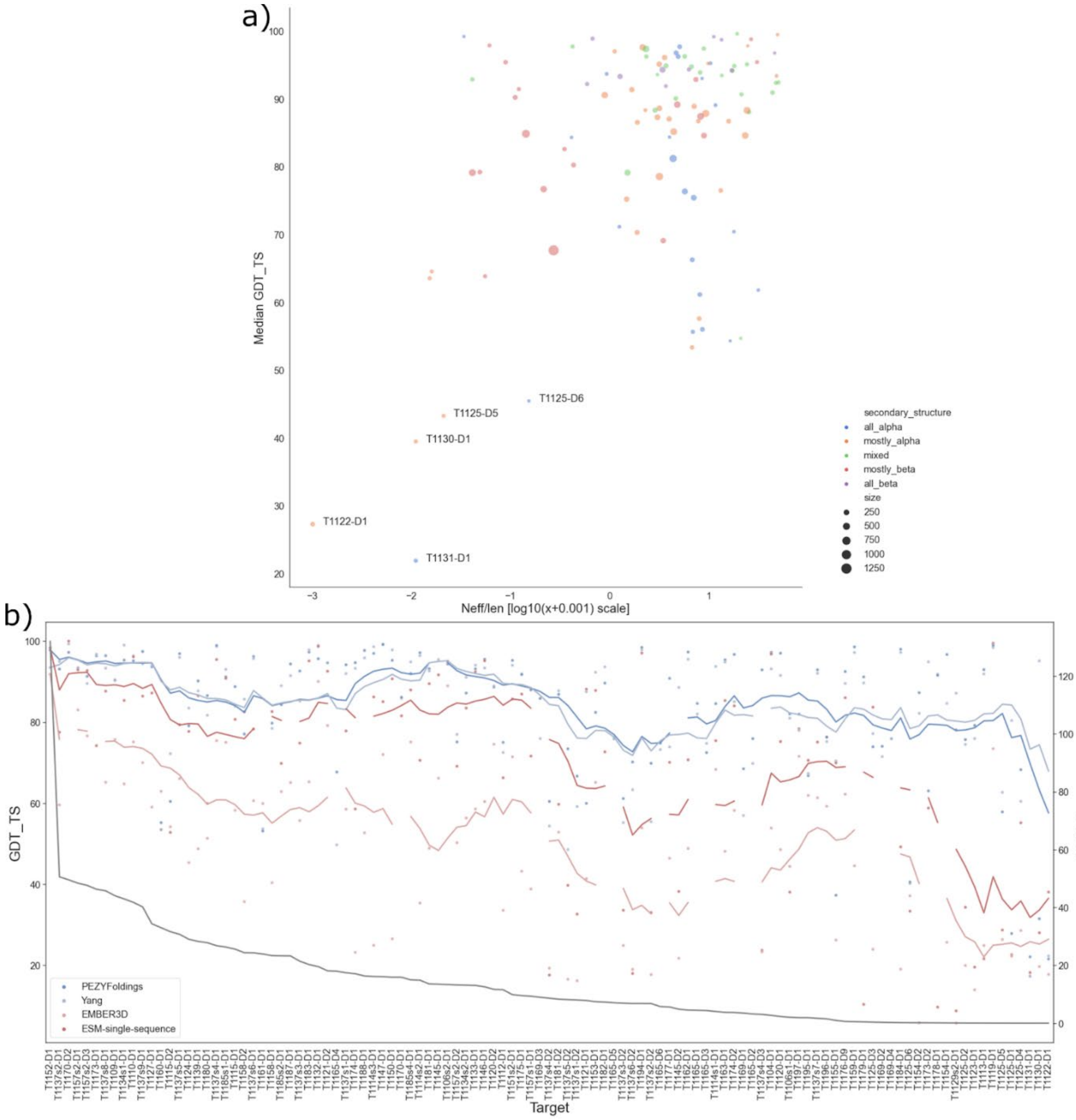


Figure 3. Analysis of performance versus Neff/length and other characteristics (a) Scatter plot of GDT_TS versus the log10 of target Neff/length + 0.001 for the top ten groups. The scatter points are coloured by secondary structure and the size of the points correspond to the size of the target. (b) Plot of GDT_TS versus target for the top performing MSA based methods (PEZYFoldings: Dark blue, Yang: Light blue) and the pLM methods (ESM-single-sequence: Dark red, EMBER3D: Light red). The lines represent a moving average for each method calculated across a ten target window. The targets are ordered by Neff/length descending from left to right with Neff/len indicated on the right-hand y-axis.

3.3.2 Examples of difficult targets

Target T1169, mosquito salivary protein SGS1, was the longest single chain target yet seen at CASP and one of the longest single chains in the PDB. It is a hard modelling target that was divided into four assessment units, three FM and one TBM-Hard [11]. Despite this splitting, indicating that most groups struggled to predict the relative orientation of the individual units, some groups produced remarkably accurate predictions of most or even the entire structure. Fig 4a shows the top-ranked prediction from the Yang-server group, with a GDT_TS of 58 overall. With the exception of the C-terminal 200 residues, the individual domains are accurately folded and packed against each other with truly impressive accuracy.

Targets T1130 and T1131 are aphid proteins which are, according to the submitting group, thought to be distantly homologous. They are both small (159 and 172 residues, respectively), largely α-helical, monomeric and were singletons in the main sequence databases at the time of CASP. A handful of good quality models of T1130 were submitted by groups who apparently discovered additional homologous sequences in the Supplementary Material of a paper [51] or in databases not searched by other groups.

Another small (241 residue) largely α-helical monomeric target that proved difficult was T1122, a cranefly nudivirus protein with unusual properties related by the experimentalist submitters: derived from viral polyhedra, crystals of T1122 obtained in the 1950s have proved stable to this day, yet dissolving the crystal denatures the protein. Its crystal structure is also unusual in containing only 25% solvent leading to a densely packed protein array. Notably the N-terminal 30 residues do not pack against the core of their own subunit, instead contacting four symmetry mates. Given its low Neff/length and the absence of this structural context as well as other contacting lattice mates numbering no fewer than 16, it is perhaps unsurprising that the best model_1, from the QUIC group, achieves a GDT_TS no better than 39 and only really authentically captures the packing of the core helices (Fig 4b).

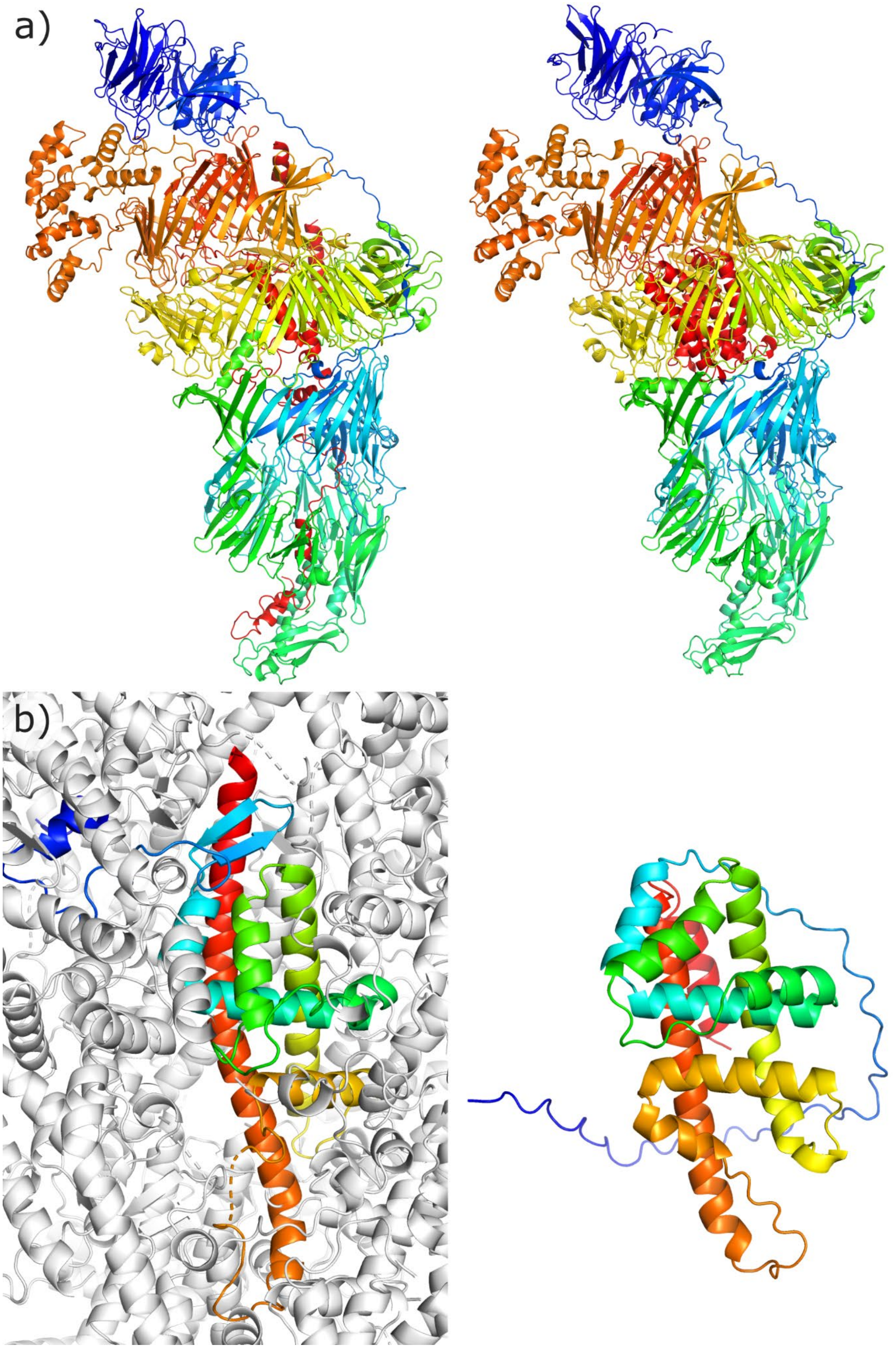


Figure 4. Examples of the best predictions produced for different targets. In each case the experimental structure is shown on the left and the prediction on the right, each coloured from blue to red from the N- to the C-terminus. a) T1169, at 2735 residues, is modelled with impressive overall accuracy by the Yang-server group, with the exception of the C-terminal 200 residues. b) the T1122 crystal structure with only 25% solvent content has a tightly packed lattice producing abundant contacts between one subunit and its neighbours (symmetry mates are shown in grey), likely contributing to the poor quality of the best prediction (from the QUIC group).

3.4 Self-assessment of results

CASP participants were asked to provide self-assessment of their models at two levels: first, by supplying local per-atom confidence values (expressed as pLDDT) in the B-factor column of their submissions, and second, by ranking their submitted models (out of five allowed per target) in order of confidence. The local confidence estimates are particularly important since it is routine to trim off lower-confidence regions for many applications of these models such as Molecular Replacement and structure searches of databases, while the model ranking is important as most user attention is likely to be focussed on the top-ranked prediction.

Detailed results of the local self-assessment are provided elsewhere in this issue [52]. Here we provide a summary of the relative group performance according to their median ASE z-scores across all submissions (Fig 5a). Notably, the confidence estimates produced by the control ColabFold and DeepMind AF2 entries are among the most accurate currently available, which should be seen as reassuring for users of the ColabFold pages [38] and the AlphaFold Protein Structure DataBase [53], respectively. In comparison, the ESMFold quality estimates are a little less accurate, something users of the new ESM Metagenomic Atlas should bear in mind [43].

Analysing model ranking data, Fig 5b shows that only around two thirds of groups out-perform the randomly expected 20% threshold of model_1 being the best submission. (Note that alternate conformation targets, where numbering of models is irrelevant for the ranking purposes, were not considered here). For example, the overall winning group PEZYFoldings and the ColabFold group are each a little below 20%. However, AF2-based methods such as ColabFold often produce several predictions which are very similar: it is clearly a greater challenge to select the best in this situation than to spot the best from among five very divergent predictions.

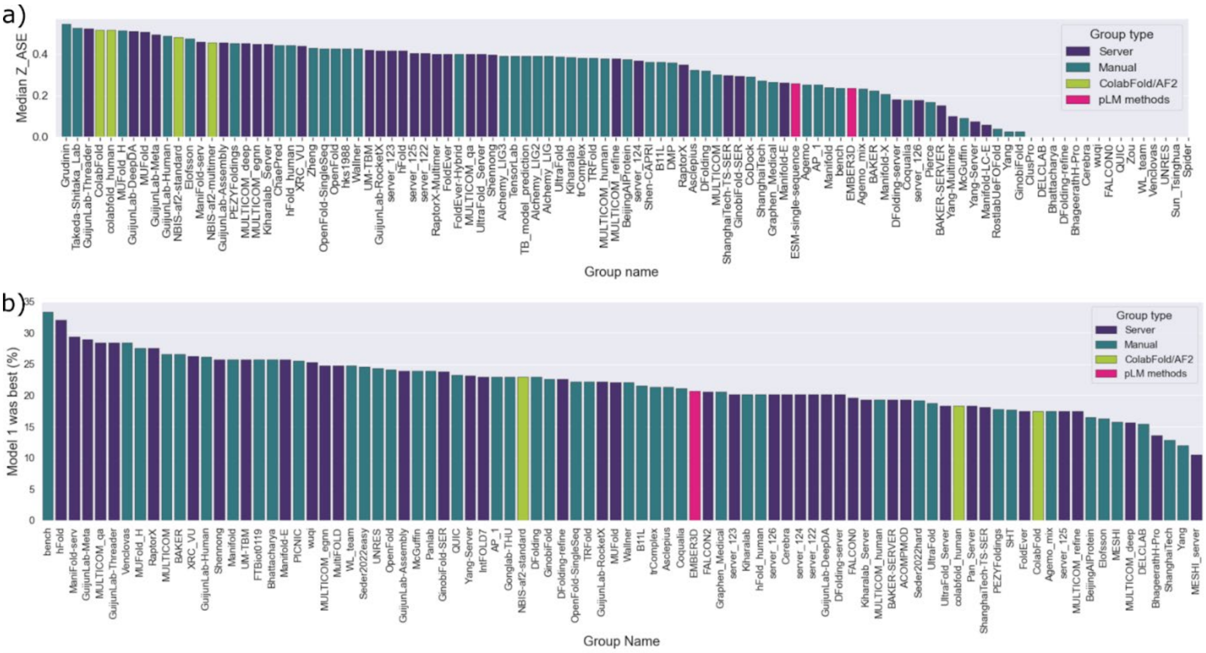


Figure 5. Group self-assessment of results (a) groups ranked by median Z_ASE. (b) groups ranked by how often model 1 was the best model (expressed as a percentage). This ranking excludes groups which attempted less than half of the targets.

3.5 Factors affecting local accuracy

Where a portion of a target structure does not necessarily capture the only accessible conformation then it is unreasonable to expect the prediction to necessarily resemble the target. Flexible surface loops are likely to be difficult to predict since there can be multiple conformations differing little energetically. Some loops will be tethered by interactions in a crystal structure, but others will retain flexibility in the context of the crystal lattice resulting in locally smeared out electron density and higher local B-factors. For four selected Deep Learning methods and the pLM method ESM-single-sequence, the relationship between experimental B-factors and residue LGA error was studied (Fig 6A). Only higher-quality (GDT_TS >80) model_1 submissions were considered (numbering from 30 in the case of ESM-single-seq to 46 for PEZYFoldings and 47 in the cases of UM-TBM, DFolding and Yang-Server). Binning residues in these predictions by normalised B-factors (see Methods) and assessing the residue LGA error range in each bin reveals a strong relationship. For all methods, mean bin error increases with increasing normalised B-factor (Fig 6A).

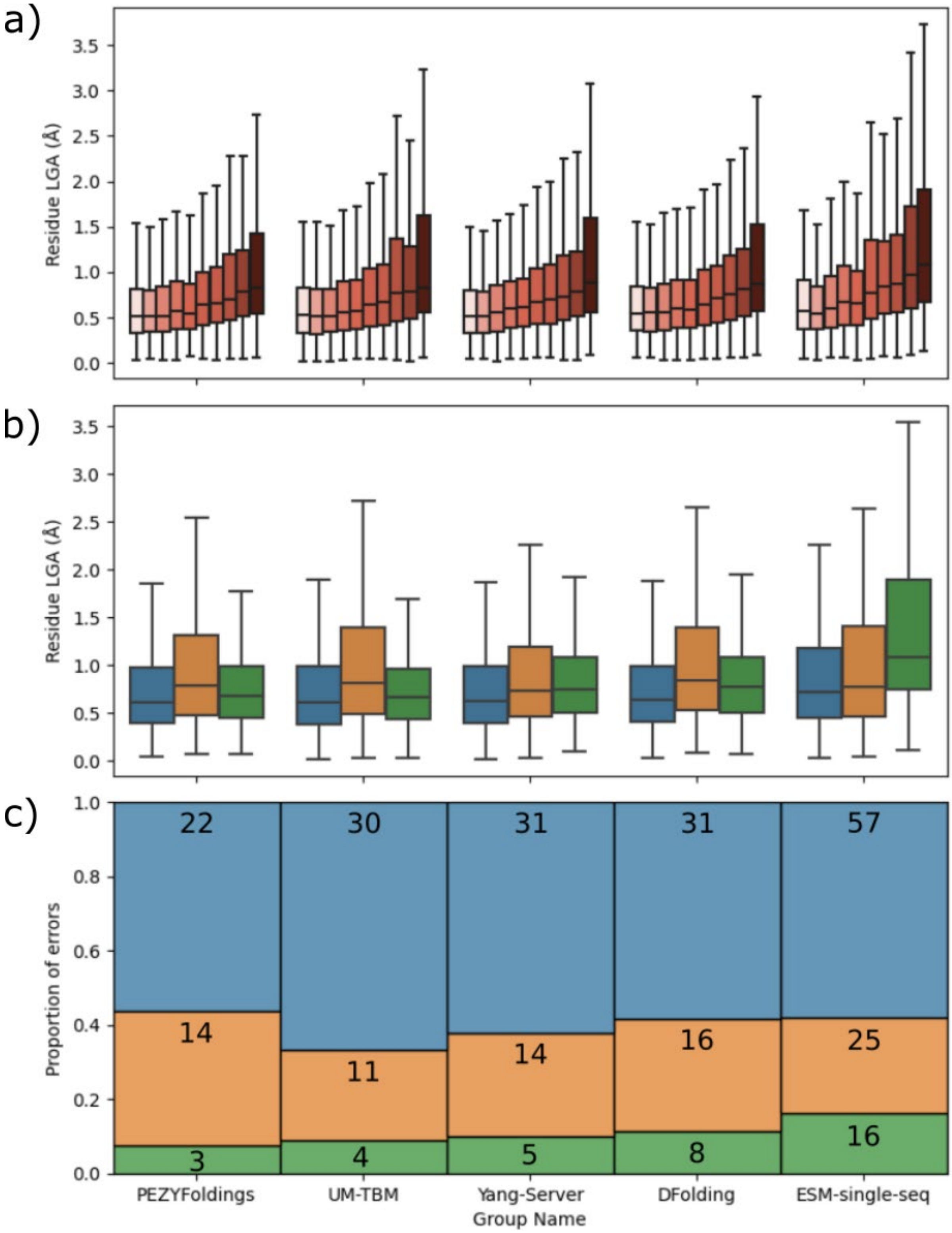


Figure 6. Factors affecting local accuracy analysed using the results of selected MSA- and pLM-based approaches. Only high-quality (GDT_TS >80) model_1 submissions are considered. (a) Residue LGA error tends to correlate with normalised B-factor: for each method, residue LGA error increases from low (light colour) to high (dark colour) bins of normalised B-factors. (b) Distribution of LGA error values across residues observed neighbouring a crystal lattice interface (orange), a chain interface (green) or neither (blue). c) Error regions (defined in Methods) are classified according to their presence at a crystal lattice interface (orange), at a chain interface (green) or neither (blue).

It is understood that the formation of the crystal lattice can lead to local distortion of residues away from the most energetically stable conformations. Indeed, it has been argued previously that where a high-quality structure prediction differs from the target at a crystal lattice, it should not automatically be inferred that the crystal structure is correct and the structure prediction in error [24]. Another way to view the situation is that the structure prediction program lacks the 3D context - the neighbouring crystal symmetry mates - that would help it accurately predict crystal lattice structures. A similar logic can be applied to interfaces between chains: lack of structural context means they may well prove harder to predict where only a single chain is being modelled. These questions were explored here first by looking at mean local error among interface residues compared to others; and secondly by defining error regions (see Methods) and checking whether there was evidence of their over-representation at crystal lattice or chain interfaces. Analysis of domain interfaces was also attempted (not shown) but the number of such residues was too small to allow meaningful analysis.

Considering first the local LGA errors (Fig 6b), residues at crystal lattice interfaces have significantly higher errors than non-interface residues for the results of all five groups considered: two sample t-test results gave p-values running from 2.73×10^{-13} to 0.04. Although the difference was often less pronounced, localization at a chain interface also resulted in significantly higher local errors in the results of all methods except PEZYfoldings. Similar results were obtained when considering error regions (see Methods) of at least three residues (Fig 6c). As expected, the pLM method ESM-single-sequence that is less accurate overall had larger numbers of such regions than the MSA-based methods. However, across all methods, a consistent proportion of around 40% of error regions are found at crystal lattice or chain interfaces, with the former always significantly outnumbering the latter. Taken together these results show that even the best predictive methods can still struggle with interface regions, especially crystal lattice contacts. However, the question remains, remembering the local forces exerted on proteins as they crystallise, as to whether crystal lattice ‘mispredictions’ should be regarded as errors: the model region may represent an alternative correct, biologically accessible conformation or even, if the lattice interface is distorted, the single biologically relevant conformation. A similar situation applies for high B-factor flexible loops where model and target, though different, may be equally valid snapshots from a biological ensemble. A continued recognition at CASP of these mitigating factors around some ‘mispredictions’ is important to define the scope for future improvement of the current state of the art.

3.6 Side chain accuracy

As explained earlier, side chain accuracy is one component of the overall composite score but, as global main chain quality has improved, particularly post-AF2, improving side chain placement is seen as an important area for future development. Assessing reasonable expectations for optimal side chain placement is complicated by the fact that surface-located side chains will often have multiple, significantly occupied conformations. These may or may not be resolved experimentally, depending largely on the resolution of the data available.

To assess if surface side chains were more difficult to model than non-surface side chains, the SCWRL4 AAA sidechain scores [21] were calculated for different types of the side chains (see Methods) (Fig 7). Seven targets were classified as all-surface (T1106s1-D1, T1114s1-D1, T1115-D2, T1119-D1, T1137s1-D2, T1137s3-D2 & T1160-D1). For the remainder, the non-surface sidechain score was higher than the surface score in all but seven targets (T1173-D1, T1137s1-D1, T1137s4-D2, T1137s4-D3, T1137s5-D2, T1137s6-D1, T1169-D1) out of the 109 targets. Remarkably, for two targets (T1161-D1; TBM-easy, and especially T1137s2-D2; FM) all non-surface side chains were correct . Fig 7 shows that side chain accuracy tends to decline, left to right, as the accuracy of the best model decreases. Naturally, the proportions of surface and non-surface residues vary across the set of targets: the median ratio of surface:non-surface residues was 2.2:1.

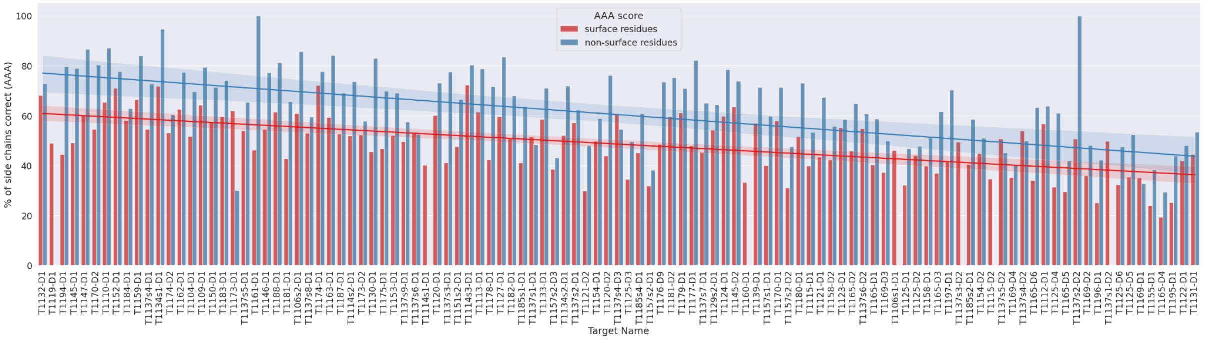


Figure 7. Per-target comparison of the mean SCWRL4 AAA sidechain score for surface residues (red) and non-surface residues (blue) for the model with the highest GDT_HA for each target. Residues were defined as surface residues if their solvent accessibility was ≥ 20% as given by the Shrake-Rupley algorithm. A line of best fit is shown for both the surface and the non-surface residues in corresponding colours. The targets are ordered in descending order by the GDT_HA value of the top model.

The side chain (SC) and backbone (BB) dihedral-based accuracy measures were plotted against each other to assess their relationship (Fig 8a). In each case, low values indicate high accuracy. The result shows a sharp dependency - as expected, high accuracy backbone structure is required before highly accurate side chain placement becomes possible [37]. Nevertheless, even with near-ideal backbones, side chain scores never approach 0 (presumably due to the alternate conformation issue mentioned above). Equally striking, high-quality backbone structures do not guarantee successful side chain placement, illustrating how these are connected but still distinct challenges for predictors.

Fig 8b shows the individual performance of the two best groups using each of the MSA-based and pLM methods. Notably, the pLM methods, especially EMBER3D, lag behind the MSA-based methods in terms of side chain prediction: for models of a given low backbone score, reflective of high quality fold prediction, pLM methods give poorer side chain placement. Figs 8c and 8d illustrate the progress from CASP14 to CASP15 in this aspect of modelling. The dramatic progress seen in Fig 8c is illustrative of the transformation brought about by AF2: the much better BB scores seen for the largely AF2-based methods at CASP15 allow, in turn, much better SC scores. Perhaps less predictably, the top two methods at CASP15, PEZYFoldings and UM-TBM, do seem to out-perform the AF2 entrant at CASP14 in terms of side chain scores (Fig 8d). Notably, the abstract from the UM-TBM team commented that its refinement element was designed specifically to improve side chain accuracy. The distribution of SC scores shows that it does slightly outperform PEZYFolding in the proportion of low SC score models.

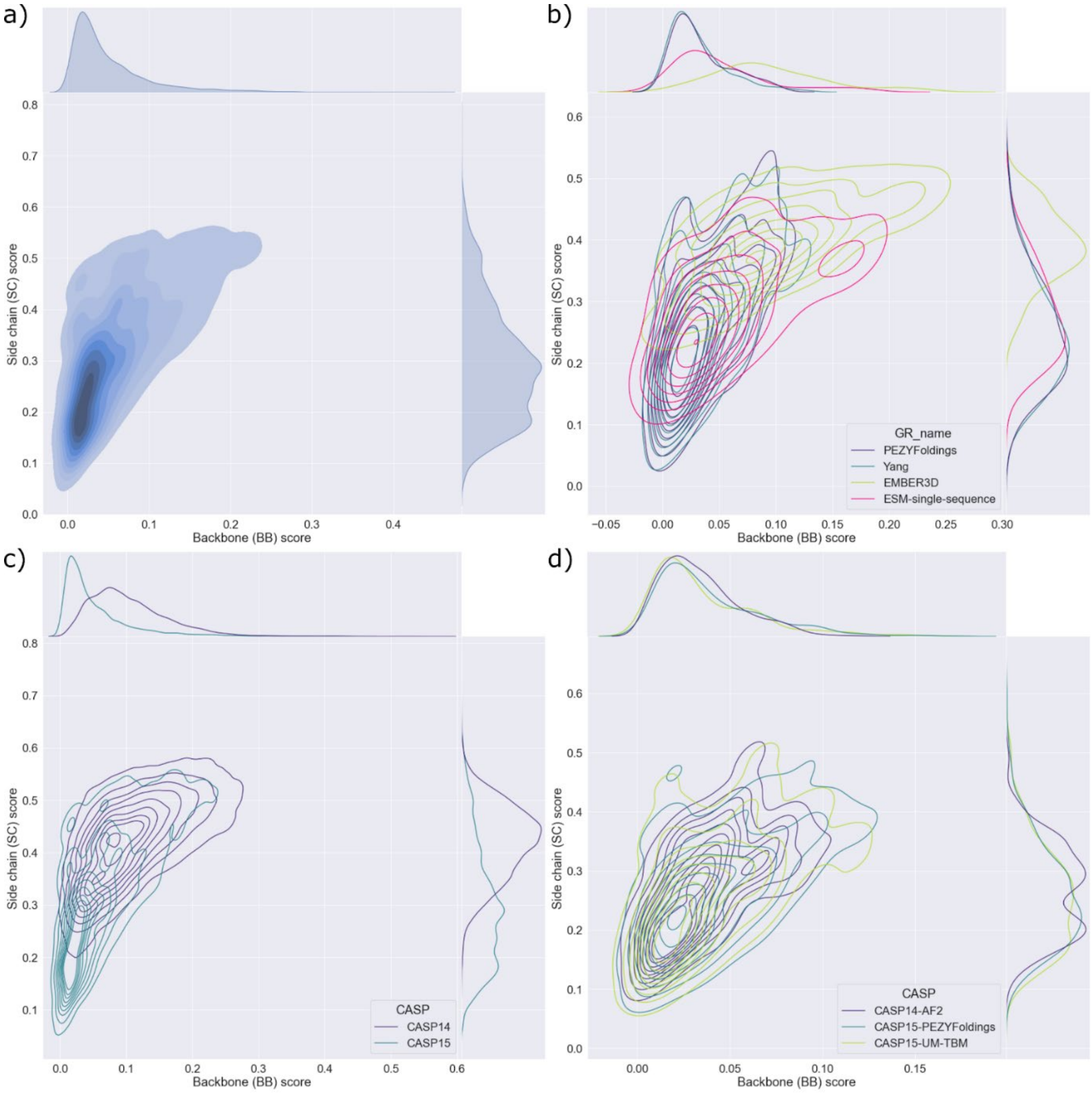


Figure 8. Contour plots illustrating the relationship between backbone and side chain dihedral scores (see Methods), calculated across whole models. Each contour plot is supplemented by a plot on the right illustrating the distribution of side chain scores, and one above showing the main chain score distribution. (a) shows all groups, all models in CASP15 illustrating how a good (low) backbone (BB) score is necessary but not sufficient for a good (low) side chain (SC) score. (b) shows a comparison between the models produced by two of the best MSA-based methods - PEZYFoldings and Yang, and two pLM methods – EMBER3D and ESM-single-sequence. (c) shows a comparison between all groups, all models in CASP14 (indigo) and CASP15 (teal). (d) shows a comparison between AF2 in CASP14 and the two top performing methods in CASP15 (PEZYFoldings and UM-TBM).

3.7 Molecular Replacement

Since Molecular Replacement (MR) is an important downstream application of protein structure modelling [20,54], submissions were also assessed directly for their suitability to serve as MR search models. As mentioned above, the reLLG [20] provides a coordinates-only metric for this purpose, and was newly included in the overall CASP15 score. However, where diffraction data were available (in 17 cases - see Supp Table 1) they were used for a more direct assessment of the CASP15 submissions. This was done first by calculating log-likelihood-gain values (LLGs) for models ideally placed by superposition on the target crystal structure and refined; and secondly by carrying out full MR using CASP submissions as search models.

3.7.1 Assessing the models' potential for success in Molecular Replacement

Submissions were processed to remove low confidence (pLDDT <70) regions, fit onto the target structure with Gesamt [28], and finally refined with Phaser [29] (see Methods for details). In this way, an LLG was obtained for model_1 submissions for each target and each group, allowing ranking of groups for each target. Converting the ranking into a score (see Methods), allowed for a comparison of groups across all targets (Fig 9a). The best scoring group was Colabfold_human and, in general, the best groups were those using AF2. Search models from pLM methods scored less well but proved to be good enough to exceed the LLG=60 threshold for many of the targets, indicating their potential utility in solving the MR problem.

In MR, it is common to split a potential search model and place the resulting domains separately. This addresses the possibility that the domain orientation in the target may be different to that in the available search model(s), whether as a result of inaccurate structure prediction or simply because of different biologically relevant inter-domain orientations. For each target except T1122 and T1125, several of the search models created from the unsplit CASP submission already scored well enough (LLG >60) to indicate potential success in MR (Fig 9b). However, when the models were split into three pieces, LLG scores improved (Fig 9c), most notably for larger targets such as T1145 (636 residues), T1174 (339 residues) and T1181 (689 residues). In addition, some of the split models showed potential for structure solution of T1125, scoring an LLG better than 60.

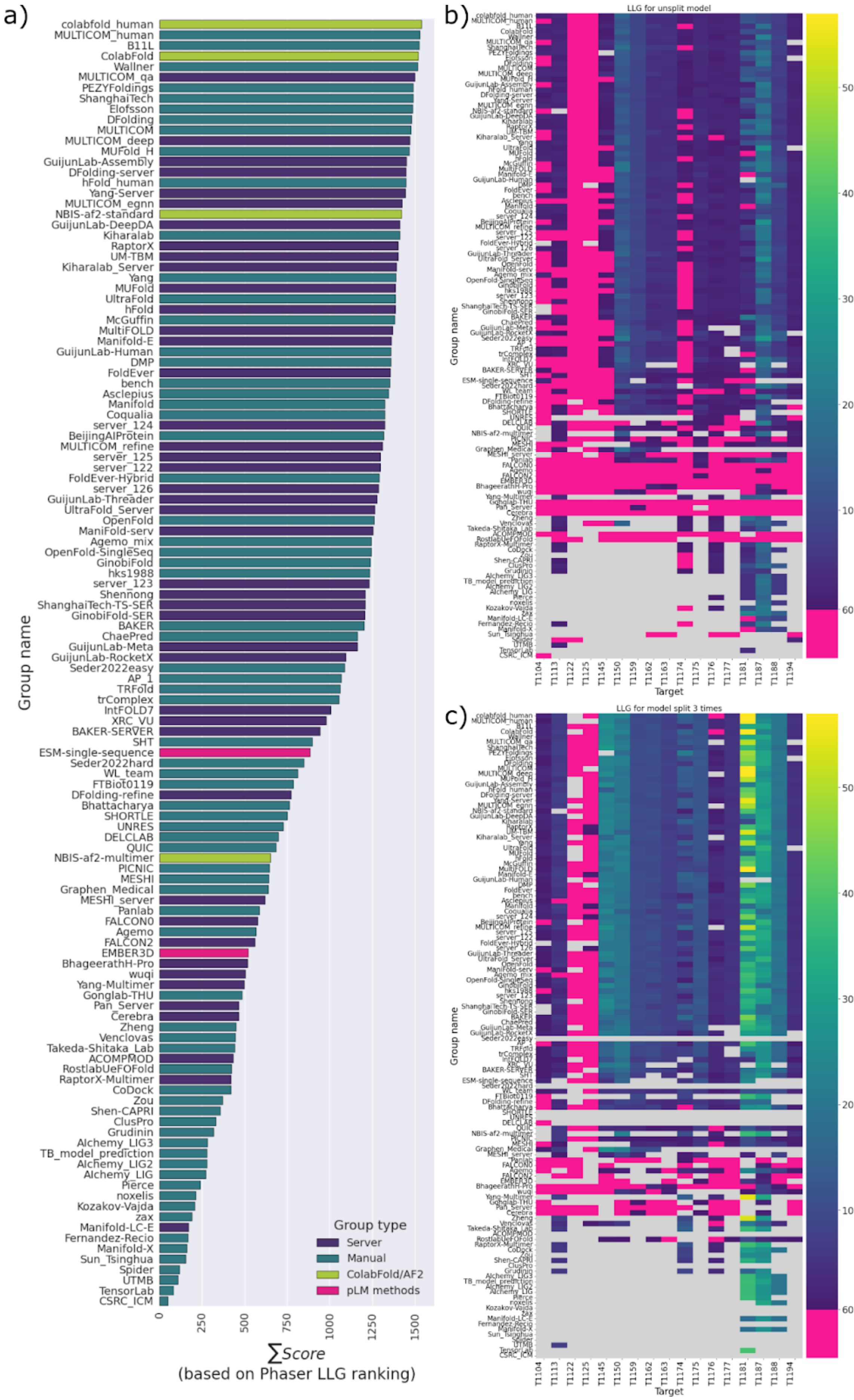

a)
colabfold_human
MULTICOM_human
B11L
ColabFold
Wallner
MULTICOM_qa
PEZYFoldings
ShanghaiTech
Elofsson
DFolding
MULTICOM
MULTICOM_deep
MUFold_H
GuijunLab-Assembly
DFolding-server
hFold_human
Yang-Server
MULTICOM_egnn
NBIS-af2-standard
GuijunLab-DeepDA
Kiharalab
RaptorX
UM-TBM
Kiharalab_Server
Yang
MUFold
UltraFold
hFold
McGuffin
MultiFOLD
Manifold-E
GuijunLab-Human
DMP
FoldEver
bench
Asclepius
Manifold
Coqualia
server_124
BeijingAIProtein
MULTICOM_refine
server_125
server_122
FoldEver-Hybrid
server_126
GuijunLab-Threader
UltraFold_Server
OpenFold
ManiFold-serv
Agemo_mix
OpenFold-SingleSeq
GinobiFold
hks1988
server_123
Shennong
ShanghaiTech-TS-SER
GinobiFold-SER
BAKER
ChaePred
GuijunLab-Meta
GuijunLab-RocketX
Seder2022easy
AP_1
TRFold
trComplex
IntFOLD7
XRC_VU
BAKER-SERVER
SHT
ESM-single-sequence
Seder2022hard
WL_team
FTBiot0119
DFolding-refine
Bhattacharya
SHORTLE
UNRES
DELCLAB
QUIC
NBIS-af2-multimer
PICNIC
MESHI
Graphen_Medical
MESHI_server
Panlab
FALCON0
Agemo
FALCON2
EMBER3D
BhageerathH-Pro
wuqi
Yang-Multimer
Gonglab-THU
Pan_Server
Cerebra
Zheng
Venclovas
Takeda-Shitaka_Lab
ACOMPMOD
RostlabUeFOFold
RaptorX-Multimer
CoDock
Zou
Shen-CAPRI
ClusPro
Grudinin
Alchemy_LIG3
TB_model_prediction
Alchemy_LIG2
Alchemy_LIG
Pierce
noxelis
Kozakov-Vajda
zax
Manifold-LC-E
Fernandez-Recio
Manifold-X
Sun_Tsinghua
Spider
UTMB
TensorLab
CSRC_ICM
Group name
0
250
500
750
1000
1250
1500
∑Score
(based on Phaser LLG ranking)
Group type
Server
Manual
ColabFold/AF2
pLM methods
b)
LLG for unsplit model
Group name
Target
T1104
T1113
T1122
T1125
T1145
T1150
T1159
T1162
T1163
T1174
T1175
T1176
T1177
T1181
T1187
T1188
T1194
5000
4000
3000
2000
1000
60
c)
LLG for model split 3 times
Group name
Target
T1104
T1113
T1122
T1125
T1145
T1150
T1159
T1162
T1163
T1174
T1175
T1176
T1177
T1181
T1187
T1188
T1194
5000
4000
3000
2000
1000
60

Figure 9 (a) Groups ranked by the cumulative LLG derived ranking score described in section 2.5.1. (bc) A comparison between the LLG scores for an ideally placed model b) before splitting and c) after splitting three times using the Birch algorithm in Slice'N'Dice. Pink indicates LLG scores below 60, the success threshold in MR. The blue to yellow gradient (see the colouring map next to the graph) depicts the LLG scores greater than 60 with yellow indicating the largest LLG values. Grey denotes instances where groups did not submit models for a target or where Phaser failed to produce a solution. Groups are ordered the same in all three panels.

3.7.2 Full Molecular Replacement

Even a search model with a high LLG after ideal placement may not succeed in full MR because of issues such as packing clashes. In order to assess real-world performance, model 1 from selected best ranking groups in the alignment tests, including the best scoring pLM-based group ESM-single-sequence, was used for full MR. The unsplit models, subjected to pLDDT-to-B-factor conversion and removal of residues scoring pLDDT<70, were used. Fig 10 shows the results of these tests, displaying the LLG from Phaser for each model and each target. No group produced a successful search for targets T1122 and T1125. For the remaining targets, many of the groups produced models which could be successfully placed in the MR search. The ESM-single-sequence models were the least successful, although they were sufficient for use as search models in several cases.

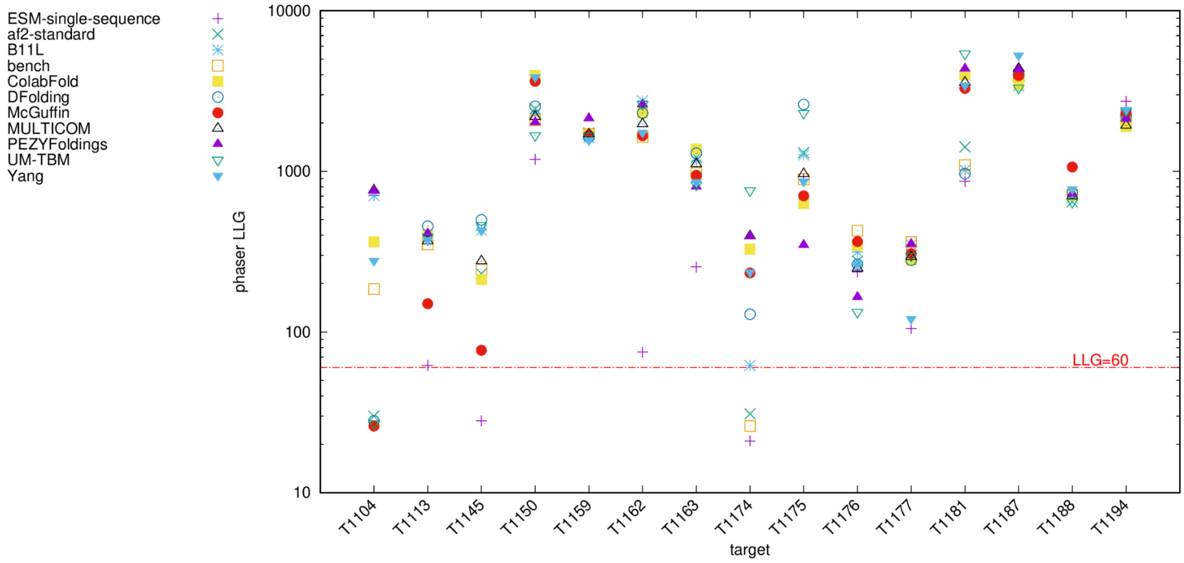
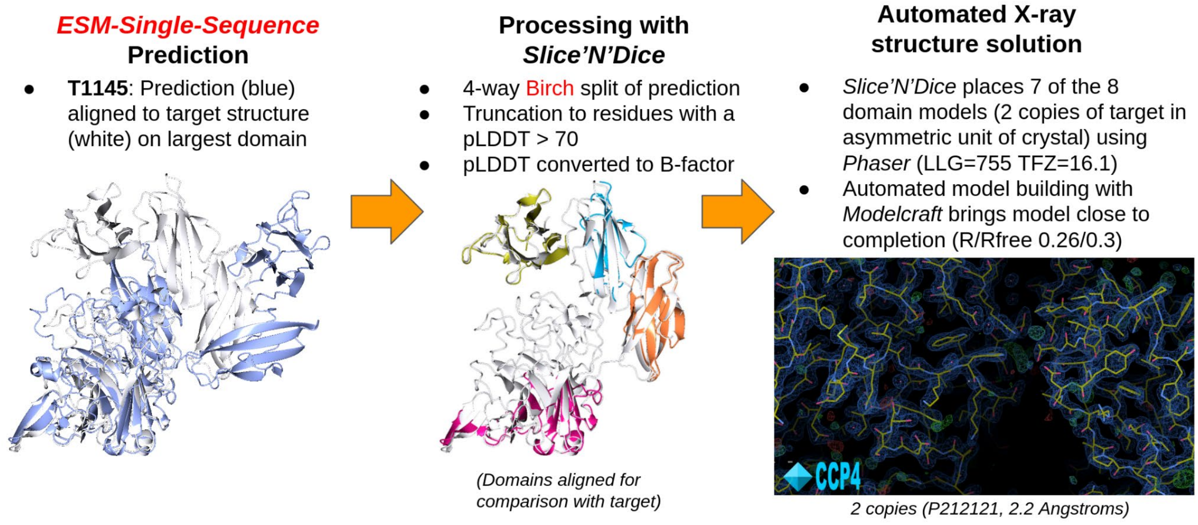


Figure 10 LLG values from full MR tests for unsplit model_1 predictions for 11 selected groups, modified to remove residues with pLDDT <70, and placed by Phaser. T1122 and T11225 are not shown since no search model produced a solution..

3.7.3 Molecular Replacement using ESM-Single-Sequence model for T1145

One of the targets where the ESM-Single-Sequence model failed to produce an MR solution was T1145. We examined this case in more detail to test whether a predicted model, which differs greatly in its overall conformation from the crystallised form, could be successfully used in a split form to produce a correct MR solution. Supp Fig 4 shows the results of the aligned model test for T1145 for model 1 from all groups. It shows that, when the ESM-Single-Sequence model is split into 2, 3 or 4 pieces using the SnD application, the resulting LLGs from Phaser strongly suggest the possibility of successful MR. Testing this hypothesis by full MR, we found that the optimal splitting was divide the predicted model into four domains (Fig 11). Target T1145 derives from a crystal structure containing two copies of the target in the asymmetric unit cell. Phaser successfully positioned seven of the eight domains producing phases and allowing calculation of an initial electron density map. The map was of sufficiently good quality for the model building application Modelcraft [34] to successfully build most of the two copies of the target structure. Restrained refinement using Refmac5 achieved an Rfactor/Rfree of 0.26/0.3 after model building, showing good agreement between the refined structure and the observed reflection data. These are typical values for what can be achieved in the automatic model building of a macromolecular crystal structure. Further completion of the structure usually requires manual effort through a graphical interface.



***ESM-Single-Sequence* Prediction**

- **T1145**: Prediction (blue) aligned to target structure (white) on largest domain

Processing with *Slice'N'Dice*

- 4-way Birch split of prediction
- Truncation to residues with a pLDDT > 70
- pLDDT converted to B-factor

(Domains aligned for comparison with target)

Automated X-ray structure solution

- *Slice'N'Dice* places 7 of the 8 domain models (2 copies of target in asymmetric unit of crystal) using *Phaser* (LLG=755 TFZ=16.1)
- Automated model building with *Modelcraft* brings model close to completion (R/Rfree 0.26/0.3)

2 copies (P212121, 2.2 Angstroms)

Figure 11. Using Slice'N'Dice to automatically split the ESM-Single-Sequence predicted model_1 for T1145 into four domains (different colours; also retaining only residues with pLDDT>70) and perform Molecular replacement using Phaser. Phaser places seven of the eight domain models and further completion is achieved using the Modelcraft model building application.

3.8 Function prediction

The interpretation and prediction of the functions of proteins based on their modelling is of major importance [55]. Some function predictions rely on global properties, such as inference of nucleic acid binding capability based on electrostatic properties [35], and are therefore tolerant of some error. Other methods are acutely dependent on the accurate capturing of fine details such as the local conformations of specific residues responsible for ligand recognition. With this in mind, targets were selected based on their interpretability by structure-based methods. Target selection was based on information given to the CASP predictors in combination with analysis and literature review of the targets. Four enzymes T1146, T1110, T1127 and T1188 with catalytic dyads or triads were selected as well as one DNA binding protein (T1151).

The ability to detect the catalytic sites by matching 3D structural motifs depends on their accurate local modelling. To assess their predictability given a certain accuracy of global modelling, the value deriving from all atom fitting of the catalytic residues of the active sites to templates was plotted against measures of global fold quality and side chain metrics. The global accuracy metric (GDT_HA) and side chain metrics (data not shown) are not strongly correlated with the RMSD of the catalytic residues, and there are outliers. Fig 12a shows this trend for the T1146 target; other targets are shown in Supp Fig 5. For example, model 1 from the QUIC group for the T1146 target has a high GDT_HA (79.4) but the RMSD for the catalytic triad residues is relatively high (2.36 Å). Inspection of this model reveals that the protein has a very accurate fold but one of the catalytic residues (His255) is in the wrong conformation (Fig 12c). Conversely, model 1 from the Agemo group has a low GDT_HA (51.3) but the RMSD (0.34 Å) for the catalytic triad residues is relatively low. Inspection of the model demonstrated that the overall fold is poorly modelled in respect of its relative domain orientation, yet the catalytic triad residues are correctly placed relative to each other (Fig 12c).

For T1151, all models score above the DNABIND default threshold (0.531) and are predicted as DNA-binding. However not all models are predicted as potentially DNA-binding by BindUP (Fig 12b). In this regard, overall model quality seems to have little impact on DNA binding prediction; there is no general trend between GDT_HA and the DNABIND probability score, nor is there a tendency for better models to predict as DNA-binding by BindUP. These findings are in line with the known error tolerance of the DNABIND method [35].

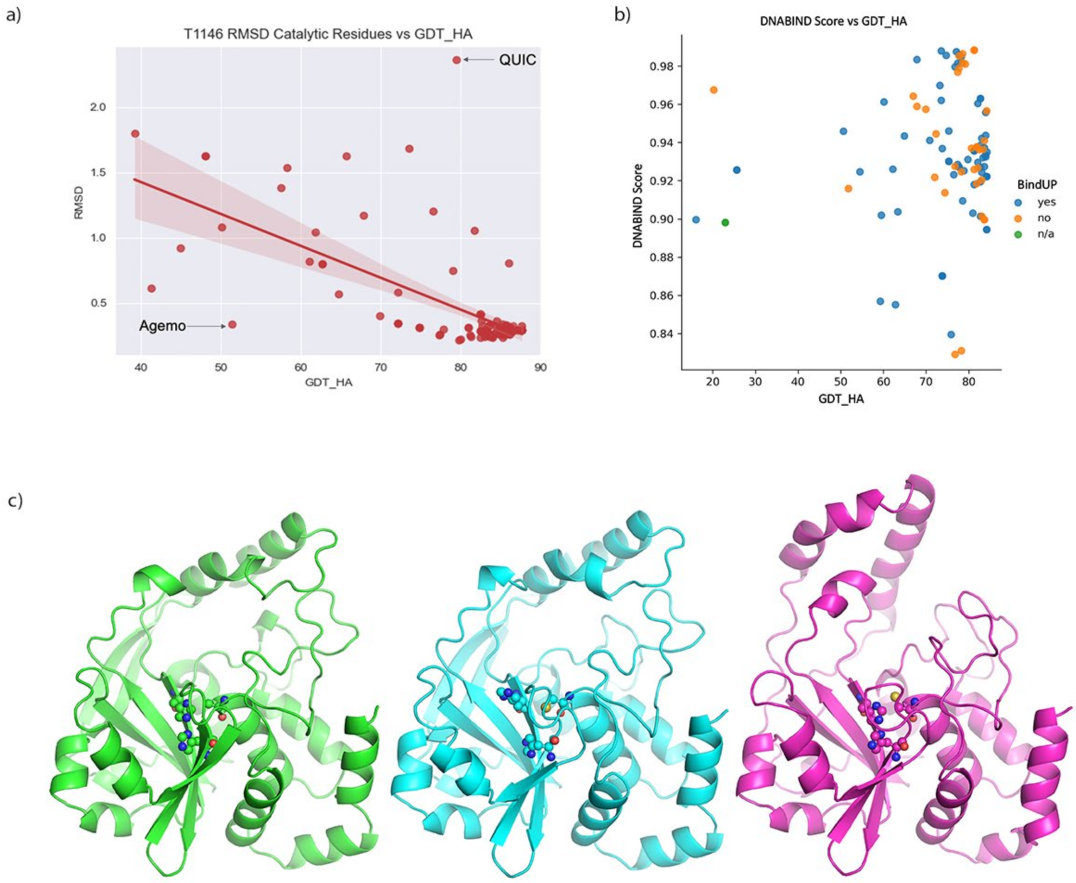


Figure 12. Function prediction based on submitted models. (a) Global accuracy and the accuracy of functional features are only weakly correlated, as exemplified here by the RMSD on catalytic residues vs GDT_HA for T1146 (b) DNABIND probability score against Global accuracy. The colouring of the data points indicates BindUP results: blue-positive, orange-negative, green-could not be processed

(c) The T1146 catalytic triad (sticks) and overall fold (cartoon) in the experimental structure (green) and two outliers; cyan - highly accurate fold, wrong conformation of one catalytic His (model 1 for the QUIC group); magenta - fold prediction less accurate, but catalytic site well-modelled (model 1 for the Agemo group).

4 | Conclusions

The first conclusion of single chain assessment at CASP15 must be that AlphaFold 2 remains the dominant presence in the field: the best-ranking group that did not use it in one form or another was the BAKER group in 28th position in the overall rankings. Set against this, however, must be placed the observation that groups used AF2 in diverse ways including, intriguingly, in combination with previous generations of structure prediction software. Thus, the UM-TBM group hybridised AF2 predictions with the I-TASSER framework [40] and the Yang-Server group sampled from both AF2 and trRosettaX2 [42]. Nevertheless, the overall winner PEZYFoldings used AF2 in relatively orthodox fashion for construction of models: a novel, additional post-prediction refinement step based on a fine-tuned AF2 turned out, on closer examination, to only modestly improve their already excellent models [44].

At CASP14 AF2 was far ahead of other groups [7,8]. This time, a large number of groups, mostly using AF2 results in one fashion or another, produced excellent models for most targets (Fig 2). What differentiated the best groups was their ability to produce good models for the most difficult set of Evaluation Units (EUs). Compared to previous CASPs there were more FM targets whose absence of templates obviously provides a first element of difficulty. Accordingly, further analysis shows that the most obvious characteristic shared by the hardest targets was a lack of detectable homologous sequences, leading to shallow MSAs and weak or even absent covariance information. This likely relates to the over-representation of fast-evolving viral sequences in the set of hardest targets. Inferred distance information from covariance analysis is known to be crucial for the initial model estimation by AF2 where templates are not available [37]. The advantages gleaned by the best-performing groups seem to partly derive from an ability to scrape hidden sequence information from sources not necessarily included in the main databases. In an ideal world, a single, unified and comprehensive database would be available to all groups so that performance disparities could be related more directly to differences in predictive methods. Aside from the number of homologues, there are hints that harder targets may be more likely to be smaller and predominantly α-helical in secondary structure. This tentative observation is in complete contrast to results in the time of fragment assembly *ab initio* methods eg [58] when these targets were generally the most favourable and is worthy of further study on a larger scale.

Even models that are globally accurate can contain regions that match the target less well. These may result straightforwardly from poorer predictive performance but other explanations are possible, and analysis supports the relevance of two other factors. First, proteins are naturally flexible and such motions can occur in the particles analysed by cryo-electron microscopy and even in a protein crystal. Such mobile regions have a variety of accessible conformations meaning that a failure of the modelling to capture the same local structure seen in the target is not necessarily indicative of erroneous modelling. In-crystal movement can result in elevated B-factors for the affected part(s) or even, in extreme cases, the complete absence of electron density. The clear trend found between higher B-factors and higher local errors (Fig 6a) suggests that in some cases the difference between prediction and target cannot necessarily be straightforwardly inferred as wrong, potentially instead being a different valid structure.

A second feature affecting interpretation of local accuracy is positioning at an interface, either at a crystal lattice interface or with a different chain in an oligomeric structure (Fig 6b). Since lattice formation can distort lattice contacts away from conformations accessible in solution, a deviation of a prediction from target at the interface may be an alternative valid structure, or even conceivably more correct than a potentially distorted conformation in the target [24]. Inter-chain interface regions are likely to be more difficult to predict than other parts of the target because their local conformation may depend on 3D structural context that is absent during the modelling process.

While analysis has focussed on difficulties and room for future improvement, it should be remembered that the overall picture is one in which many groups produce remarkably good models for most targets (Fig 2, Table 1). Furthermore, as also noted elsewhere [20,54], outputs from readily available methods like ColabFold [38] and ESMFold [43] (or deposits in databases generated by similar protocols [43,53]) can solve most crystal structures by Molecular Replacement (Figs 10, 11); are similarly valuable in solution of structures determined by cryo-Electron Microscopy as starting models for density-guided refinement into experimental data [56]; and often allow functional annotation by accurately capturing key local features (Fig 12). In this respect, much credit goes to DeepMind for making AF2 Open Access and thereby democratising state-of-the art-modelling and reinvigorating whole areas of research. Notably, since human groups have typically out-performed servers at previous competitions, and recalling that the stand-out CASP14 winner DeepMind competed as a human group, the strong representation of automated servers among the very best groups at CASP15 is a welcome development (Fig 1, Supp Figs 1 and 2). In summary, while further methods development will proceed apace, addressing issues such as side chain accuracy [57] and targets with few homologues, colleagues across biology already have immensely powerful tools whose applications will only continue to expand.

ACKNOWLEDGEMENTS

This research was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) grant BB/S007105/1 (DJR), US National Institute of General Medical Sciences (NIGMS/NIH) grant R01GM100482 (AK). The authors thank Claudio Mirabello at the National Bioinformatics Infrastructure Sweden at SciLifeLab for providing baseline NBIS-AF2-standard models with their MSA and Neff parameters. Computation of the baseline models was performed at NSC Tetralith/BerzeLiUs provided by the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

Tables

Table 1.

Overall quality metrics for model_1 submissions by selected methods. The best-performing MSA-based methods, the AF2 controls and the best pLM-based method were chosen for the comparison. The GDT_TS threshold of 45 corresponds to a prediction of the correct topology [8]; the 90 threshold is taken as the approximate difference between two crystal structures of the same protein [7].

Method	GDT_TS ≥ 45 (total targets)	GDT_TS ≥ 90 (total targets)	Median GDT_TS
PEZYFoldings	101 (107) = 94.4%	53 (107) = 49.5%	89.65
UM-TBM	105 (109) = 96.3%	46 (109) = 42.2%	87.36
Yang	105 (108) = 97.2%	51 (108) = 47.2%	89.26
ColabFold	93 (109) = 85.3%	42 (109) = 38.5%	86.67
NBIS-af2-standard	93 (109) = 85.3%	38 (109) = 34.9%	85.88
ESM-single-sequence	72 (93) = 77.4%	19 (93) = 20.4%	77.71

References

[1] Moult J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. *Proteins.* 1995;23(3):ii - v.

[2] 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):1607-1617.

[3] Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res*. 2021;49(D1):D437-D451.

[4] Defay T, Cohen FE. Evaluation of current techniques for ab initio protein structure prediction. *Proteins*. 1995;23(3):431-445.

[5] Leman JK, Weitzner BD, Lewis SM, Adolf-Bryfogle J, Alam N, Alford RF, et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods*. 2020;17(7):665-680.

[6] AlQuraishi M. Machine learning in protein structure prediction. *Curr Opin Chem Biol*. 2021;65:1-8.

[7] Pereira J, Simpkin AJ, Hartmann MD, Rigden DJ, Keegan RM, Lupas AN. High-accuracy protein structure prediction in CASP14. *Proteins*. 2021;89(12):1687-1699.

[8] Kinch LN, Pei J, Kryshtafovych A, Schaeffer RD, Grishin NV. Topology evaluation of models for difficult targets in the 14th round of the critical assessment of protein structure prediction (CASP14). *Proteins*. 2021;89(12):1673-1686.

[9] Chowdhury R, Bouatta N, Biswas S, Floristean C, Kharkar A, Roy K, et al. Single-sequence protein structure prediction using a language model and deep learning. *Nat Biotechnol*. 2022;40(11):1617-1623.

[10] Wang W, Peng Z, Yang J. Single-sequence protein structure prediction using supervised transformer protein language models. *Nat Comput Sci*. 2022;2(12):804-814.

[11] Kryshtafovych A, Rigden D. To split or not to split: CASP15 targets and their processing into tertiary structure evaluation units. Published online March 13, 2023. doi:10.22541/au.167872023.39044035/v1

[12] 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.

[13] Olechnovič K, Kulberkytė E, Venclovas C. CAD-score: a new contact area difference-based function for evaluation of protein structural models. *Proteins*. 2013;81(1):149-162.

[14] Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. *Proteins*. 2014;82 Suppl 2(0 2):7-13.

[15] Croll TI, Sammito MD, Kryshtafovych A, Read RJ. Evaluation of template-based modeling in CASP13. *Proteins*. 2019;87(12):1113-1127.

[16] Chen VB, Arendall WB 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr*. 2010;66(Pt 1):12-21.

[17] Pereira J, Lamzin VS. A distance geometry-based description and validation of protein main-chain conformation. *IUCrJ*. 2017;4(Pt 5):657-670.

[18] Zemla A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res*. 2003;31(13):3370-3374.

[19] Kryshtafovych A, Monastyrskyy B, Fidelis K, Moult J, Schwede T, Tramontano A. Evaluation of the template-based modeling in CASP12. *Proteins*. 2018;86 Suppl 1(Suppl 1):321-334.

[20] Millán C, Keegan RM, Pereira J, Sammito MD, Simpkin AJ, McCoy AJ, et al. Assessing the utility of CASP14 models for molecular replacement. *Proteins*. 2021;89(12):1752-1769.

[21] Krivov GG, Shapovalov MV, Dunbrack RL Jr. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*. 2009;77(4):778-795.

[22] Zhou Y, Kloczkowski A, Faraggi E, Yang Y. *Prediction of Protein Secondary Structure*. Humana; 2016.

[23] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983;22(12):2577-2637.

[24] Simpkin AJ, Sánchez Rodríguez F, Mesdaghi S, Kryshtafovych A, Rigden DJ. Evaluation of model refinement in CASP14. *Proteins*. 2021;89(12):1852-1869.

[25] Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J Mol Biol*. 1973;79(2):351-371.

[26] SelectBySASA. https://www.rosettacommons.org/docs/latest/scripting_documentation/RosettaScripts/TaskOperations/taskoperations_pages/SelectBySASAOperation. Accessed May 9, 2023

[27] Simpkin AJ, Elliott LG, Stevenson K, Krissinel E, Rigden D, Keegan RM. Slice'N'Dice: Maximising the value of predicted models for structural biologists. *bioRxiv*. Published online July 2, 2022. doi:10.1101/2022.06.30.497974

[28] Krissinel E. Enhanced fold recognition using efficient short fragment clustering. *J Mol Biochem*. 2012;1(2):76-85.

[29] Read RJ. Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Crystallogr D Biol Crystallogr*. 2001;57(Pt 10):1373-1382.

[30] McCoy AJ, Oeffner RD, Wrobel AG, Ojala JRM, Tryggvason K, Lohkamp B, et al. Ab initio solution of macromolecular crystal structures without direct methods. *Proc Natl Acad Sci U S A*. 2017;114(14):3637-3641.

[31] GitHub - cctbx/cctbx_project: Computational Crystallography Toolbox. GitHub. https://github.com/cctbx/cctbx_project. Accessed May 10, 2023

[32] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.

[33] Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr*. 1997;53(Pt 3):240-255.

[34] Bond PS, Cowtan KD. ModelCraft: an advanced automated model-building pipeline using Buccaneer. *Acta Crystallogr D Struct Biol*. 2022;78(Pt 9):1090-1098.

[35] Szilágyi A, Skolnick J. Efficient prediction of nucleic acid binding function from low-resolution protein structures. *J Mol Biol*. 2006;358(3):922-933.

[36] Paz I, Kligun E, Bengad B, Mandel-Gutfreund Y. BindUP: a web server for non-homology-based prediction of DNA and RNA binding proteins. *Nucleic Acids Res*. 2016;44(W1):W568-W574.

[37] 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):583-589.

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

[39] 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):871-876.

[40] Zheng W, Wuyun Q, Freddolino PL, Zhang Y. Integrating deep learning with multi-MSA and threading alignments for high quality protein monomer and complex structure prediction in CASP15. *Proteins*. Published online 2023.

[41] Zhou X, Zheng W, Li Y, Pearce R, Zhang C, Bell EW, et al. I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction. *Nat Protoc*. 2022;17(10):2326-2353.

[42] Peng Z, Wang W, Wei H, Yang J. Combination of trRosettaX2 and AlphaFold2 for protein structure prediction in CASP15. *Proteins*. Published online 2023.

[43] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein

structure with a language model. *Science*. 2023;379(6637):1123-1130.

[44] Oda T. Improving protein structure prediction with extended sequence similarity searches and deep-learning-based refinement in CASP15. Published online April 17, 2023. doi:10.22541/au.168170992.27078535/v1

[45] Fisher RA. On the interpretation of χ 2 from contingency tables, and the calculation of P. *J R Stat Soc*. 1922;85(1):87.

[46] Kryshtafovych A, Schwede T, Topf M, Fidelis K, J. M. Critical assessment of methods of protein structure prediction (CASP)-Round XV. *Proteins*. Published online 2023.

[47] Wu T, Hou J, Adhikari B, Cheng J. Analysis of several key factors influencing deep learning-based inter-residue contact prediction. *Bioinformatics*. 2020;36(4):1091-1098.

[48] Skwark MJ, Raimondi D, Michel M, Elofsson A. Improved contact predictions using the recognition of protein like contact patterns. *PLoS Comput Biol*. 2014;10(11):e1003889.

[49] Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin AMJJ. Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins*. 2018;86 Suppl 1(Suppl Suppl 1):51-66.

[50] Weissenow K, Heinzinger M, Rost B. Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. *Structure*. 2022;30(8):1169-1177.e4.

[51] Stern DL, Han C. Gene Structure-Based Homology Search Identifies Highly Divergent Putative Effector Gene Family. *Genome Biol Evol*. 2022;14(6). doi:10.1093/gbe/evac069

[52] Studer G, Tauriello G, Schwede T. Assessment of the assessment - All about complexes. *Proteins*. Published online 2023.

[53] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res*. 2022;50(D1):D439-D444.

[54] McCoy AJ, Sammito MD, Read RJ. Implications of AlphaFold2 for crystallographic phasing by molecular replacement. *Acta Crystallogr D Struct Biol*. 2022;78(Pt 1):1-13.

[55] Rigden DJ, ed. *From Protein Structure to Function with Bioinformatics*. 2nd ed. Springer; 2017.

[56] Mulvaney T, Elliott L, Beton J, Kretsch R, Sweeney A, Rigden D, et al. Refinement of CASP15 CryoEM Targets. *Protein Sci*. Published online 2023.

[57] Wu T, Guo Z, Cheng J. Atomic protein structure refinement using all-atom graph representations and SE(3)-equivariant graph transformer. *Bioinformatics*. Published online May 5, 2023. doi:10.1093/bioinformatics/btad298

[58] Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. *Nucleic acids research.* 2004 Jul 1;32(suppl_2):W526-31.