Received: 30 April 2019 | Revised: 26 June 2019 | Accepted: 19 July 2019

DOI: 10.1002/prot.25787

PROTEINS WILEY

RESEARCH ARTICLE

A further leap of improvement in tertiary structure prediction in CASP13 prompts new routes for future assessments

Luciano A. Abriata[1,2] | Giorgio E. Tamò[1,2] | Matteo Dal Peraro[1,2]

[1]School of Life Sciences, Institute of Bioengineering, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[2]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

Correspondence
Luciano A. Abriata and Matteo Dal Peraro, School of Life Sciences, Institute of Bioengineering, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne 1015, Switzerland.
Emails: luciano.abriata@epfl.ch (L. A. A.) and matteo.dalperaro@epfl.ch (M. D. P.)

Funding information
Swiss National Science Foundation

Abstract

We present our assessment of tertiary structure predictions for hard targets in Critical Assessment of Structure Prediction round 13 (CASP13). The analysis includes (a) assignment and discussion of best models through scores-aided visual inspection of models for each evaluation unit (EU); (b) ranking of predictors resulting from this evaluation and from global scores; and (c) evaluation of progress, state of the art, and current limitations of protein structure prediction. We witness a sizable improvement in tertiary structure prediction building on the progress observed from CASP11 to CASP12, with (a) top models reaching backbone RMSD <3 Å for several EUs of size <150 residues, contributed by many groups; (b) at least one model that roughly captures global topology for all EUs, probably unprecedented in this track of CASP; and (c) even quite good models for full, unsplit targets. Better structure predictions are brought about mainly by improved residue-residue contact predictions, and since this CASP also by distance predictions, achieved through state-of-the-art machine learning methods which also progressed to work with slightly shallower alignments compared to CASP12. As we reach a new realm of tertiary structure prediction quality, new directions are proposed and explored for future CASPs: (a) dropping splitting into EUs, (b) rethinking difficulty metrics probably in terms of contact and distance predictions, (c) assessing also side chains for models of high backbone accuracy, and (d) assessing residue-wise and possibly residue-residue quality estimates.

KEYWORDS

contact prediction, critical assessment of structure prediction, distance prediction, homology modeling, machine learning, molecular modeling, residue coevolution, sequence alignment, structural bioinformatics, structure prediction

1 | INTRODUCTION

The biannual Critical Assessment of Structure Prediction (CASP) aims to provide an objective evaluation of state-of-the-art methodologies in protein structure prediction.[1] Predictors submit models of target proteins whose structures are withheld from public release during the experiment, and teams of assessors evaluate the models along different tracks. In this article, we describe our assessment of CASP13 tertiary structure predictions for the most challenging target evaluation units (EUs). These EUs encompass 32 cases assigned by Kinch et al[2] to the Free Modeling (FM) category and 15 cases assigned to the FM/TBM category. FM EUs are those for which no clear templates were available in the Protein Data Bank (PDB) at the time of the experiment neither at the sequence nor structural levels. Notice that this leaves place for some FM EUs that do have somewhat structurally similar templates in the PDB but at extremely low or even null sequence similarity. In turn, EUs of the FM/TBM category are those for which templates were available in the PDB (hence in principle

amenable to Template-Based Modeling) but they were substantially different from the target in terms of sequence and/or some structural details. Notice that the exact classification of each EU as FM or FM/TBM depends on a number of considerations specific to each target, detailed in the article by Kinch et al. As a last word regarding the analyzed targets, in this CASP we further evaluated predictions for four full targets that contain mixed FM, FM/TBM, and TBM domains; these are flagged "FM-special" at the Prediction Center website.

We begin by describing examples of scores-aided visual inspection of models for cases representative of the kinds of situations we encountered upon evaluation, and we provide all the detailed analysis of best models submitted for all EUs in the Annex 1 of Data S1. This assessment readily highlights one outstanding group, A7D from DeepMind (043), which submitted models evaluated as best for the largest number of EUs. Then, *Z*-scores based on combination of the two most useful metrics for guiding model evaluation, Global Distance Test Total Score (GDTTS) and QCS, provide the official CASP13 ranking for the tertiary structure prediction track, where A7D ranks first followed by Zhang, MULTICOM, QUARK, and Zhang-server. This ranking is invariable to other tested scores like GDTTS-only as used in CASP12, QCS-only, or TM-only. We also highlight a few other groups who did not make it to the top five of the global ranking but submitted single best models for certain EUs. We then analyze progress in this track of CASP over time exploring potential causes and still open problems. We end up with a discussion on considerations for future CASPs regarding the definition and classification of EUs and the possibility of making more thorough evaluations as backbones get better modeled, by considering local details, side chain conformations, and quality estimates.

2 | METHODS

2.1 | Scores-guided visual evaluation and assignment of best models

Our method for model analysis is based on the strategy we employed in CASP12,[3] itself inspired from the CASP10 assessment[4] and executed through an improved version of the CASP12 web app expanded with additional metrics, plots, and data. The CASP13 web app is available at http://predictioncenter.org/casp13/casp13-topology-assessment/ with an https alternative/backup at https://lucianoabriata.altervista.org/papersdata/casp13fmassessment/.

In brief, we first clustered models at 3 Å Cα RMSD keeping those of highest GDTTS as representatives of each cluster and listing all models that belong to each cluster (GDTTS[5] reports an average of the maximum number of model residues that can be superimposed onto the target under cutoffs of 1, 2, 4, and 8 Å). The clustering procedure reduced the set of models to be evaluated by 2-fold to 5-fold in this CASP. Evaluation of cluster representatives proceeded through the interactive web app. Being totally web based,[6,7] it runs on any web browser and does not require any plugins, thus being seamlessly available to the reader (although target structures are of course made available only after release by the PDB). This web app allows interactive navigation of files containing scores and three-dimensional (3D) models for all cluster 0representatives of FM and FM/TBM EUs, individual models within clusters, and target structures if available in the PDB. All 3D views are mouse synchronized to facilitate comparison. Targets can be color c by residue index or by B-factors, while models can be color coded by residue index or by quality estimates (which the predictors are instructed to place in the B-factor column).

The scores used for short-listing models within the web app include GDTTS[5] defined earlier, plus QCS,[8] Handedness, correlation of distance matrices (CoDM) and deformation (DFM) scores (these five as in the CASP10 and CASP12 assessments[3,4]), and the TM score.[9,10] GDTTS ranges between 0 and 100 with scores below 20 indicating very poor models while a score of 100 corresponds to full match of all Cα coordinates within 1 Å deviation. GDTTS values above 50 generally indicate overall good topology,[8] while in our experience, values of 35-40 are the lower bound for some very coarse topology being captured. The QCS score[8] (defined from 0 to 100, higher is better) is based on comparison of secondary structures and was devised to mimic expert inspection. The DFM[4] score (defined positive and such that lower values are better) measures distortion over residue tetrads in the model relative to the target. The Handedness[4] score (defined from 0 to 1, higher is better) compares global conformations by looking at the relative orientations of randomly picked atom tetrads. The CoDM[4] score (defined from −1 for perfect anticorrelation to 1 for perfect correlation) measures the correlation of the residue-residue distance matrices of the target and model. Finally, the TM[9] score (defined from 0 to 1, higher is better) combines the distances between all residues of the aligned region in the target and model, weighting them by a reference distance calibrated from the target's length instead of considering only residues within selected distance cutoffs like GDTTS.[9,10] All these scores and others available in the web app were provided by the Prediction Center.[11]

When the user selects an EU from the dropdown list, the web app displays correlation plots of GDTTS vs TM, QCS, CoDM, Handedness, and DFM scores, ranks the cluster representatives according to each of these six scores, and pools them together into a list of models to be visually inspected in comparison to the target EU (5 top clusters by each metric are short-listed, giving a maximum possible number of 30 models; however, in most cases less than 30 clusters are selected because the scores are correlated). Mouse hovering over the name of each cluster's representative in this list displays a full list of all models in the cluster. The user can either inspect models and structures in 3D online through mouse-synchronized JSmol[12] applets, or download a PyMOL script for offline analysis. Online models representative of the clusters can be compared by clicking on the score correlation plots or on the list of proposed top clusters.

Compared to the CASP12 web app, the new web app includes (a) distinct coloring of human and server groups in the scores plots (blue indicates at least one model of the cluster was submitted by a server), (c) the possibility to color-code models by residue-wise quality estimates if submitted, and (c) additional plots including correlation of GDTTS against other metrics such as LDDT,[13] SphereGrinder,[14] and MolProbity,[15] as well as a correlation plot of RMSD vs coverage.

Besides, we have created web app versions for target EUs clustered at 1 Å RMSD (extended information on the web at http://predictioncenter.org/casp13/casp13-topology-assessment/index_1A.html) and for full unsplit targets (at http://predictioncenter.org/casp13/casp13-topology-assessment/index_nosplits.html). As discussed by the end of this article and as exemplified in some cases, these extensions might become useful in future CASPs to achieve deeper assessments now that models are getting increasingly good, by looking at local details, full targets, quality estimates, and so forth.

2.2 | Ranking of predictor groups

For group ranking we used *Z*-scores on equally weighed GDTTS and QCS. The *Z*-scores are initially calculated for each target from the distribution of GDTTS scores of models designated as #1; then, all the models that scored worse than two SD from the distribution mean are assigned a *Z*-score of −2.0 and removed from the model set. New *Z*-scores are then calculated on the outlier-free set. Models with a new *Z*-score less than −2.0 are also assigned a value of −2.0. The final *Z*-scores are averaged over the corresponding target sets and used to rank the performance of participating groups. Other tested scores (eg, on GDTTS alone or on QCS alone) were calculated in the same way. All *Z*-scores reported are calculated by and available at the Prediction Center.[11]

2.3 | Other methods and resources

HHpred and LGA are the main metrics employed for assessment of EU difficulty upon EU definition and classification. These two metrics are explained in detail in Reference [16] and applied to EU classification in the CASP13 article by Kinch et al in this issue. Briefly, HHpred is the product of the HHblits score for the best sequence-level match to the PDB and the sequence coverage of this hit; whereas LGA measures the structural similarity to the closest template available at the PDB regardless of sequence.

The alignment depth Neff is the ratio of the number of sequences retrieved from an HHblits run relative to the length of the query sequence in number of residues. For its calculation, the HHblits program (HHsuite3.0-beta0.3 version) was run on the uniclust30 database as of April 2018 with an *E*-value threshold of 10^{-3}, three iterations, and a minimum coverage of 60%.

Performance vs CASP round was plotted from the CASP12 plot[3] and CASP13 data. All the data presented in the web app, HHpred (computed as HHblits score × coverage) and LGA score presented here, Neff, Grishin plots, were provided by the Prediction Center.

Finally, all models for CASP13 and previous CASPs are freely available for download at the prediction center at http://predictioncenter.org/download_area/. They are also searchable through sequence queries (potentially interesting because some targets were canceled due to missing structures, but models are available at the Prediction Center) with the tool at http://lucianoabriata.altervista.org/modelsearch/.

3 | RESULTS AND DISCUSSION

Our assessment of tertiary structure predictions for hard (FM and FM/TBM) target EUs in CASP13 includes (a) assignment of best models through scores-aided visual inspection for each EU; (b) ranking of predictors resulting from this evaluation and from global scores; and (c) evaluation of progress and state of the art of protein structure prediction. Given that the rankings and the progress report foster a lengthy discussion of utmost importance to CASP, we decided to present the target-specific evaluation only briefly here, highlighting the relevance of using multiple metrics to assist expert comparison of models to targets and overviewing cases of excellent and poor predictions for a few EUs and full targets. In turn, the full evaluation of all target EUs is available in Annex 1 of Data S1.

3.1 | Scores-guided visual assignment of best predictions for each target EU and for full targets

With our web app (Figure S1, http://predictioncenter.org/casp13/casp13-topology-assessment/ or https://lucianoabriata.altervista.org/papersdata/casp13fmassessment/) we can efficiently navigate through all 3-Å clusters of models in two-dimensional plots where GDTTS is resolved against each one of QCS, CoDM, Handedness, DFM, and TM. Other scores are further available; particularly useful among these plots is that of "RMSD against fraction of aligned residues," as it is easy to grasp and very informative especially for very good and excellent models. In the web app plots and table of shortlisted clusters, by stepping on one cluster we can inspect a representative model and its target in synchronized 3D. We can thus make a score-informed but expert-based judgment about which models capture the topology and tertiary structures best for each EU, as we describe in Data S1. As targets are released by the PDB, the reader can experience on the web app the evaluation process as described. We finally note here that this year we have introduced additional web apps for EUs clustered at 1 Å and for some full targets (links available in the main web app and in Section 2).

It is critical to acknowledge that no score provides a gold standard for evaluation, and this is why we need to look at multiple scores, as we now exemplify starting from the easiest to the hardest cases to assess.

In some cases, especially when the best models are good to excellent (roughly GDTTS above 60-70), all scores tend to correlate linearly with GDTTS, with two situations showing up, as follows. One situation arises when the top cluster of models is the same by all scores and is far from the runner-up along all or many scores. In these cases, it most often happens that the top cluster is indeed the best one upon visual inspection too, as exemplified by T0955-D1 with a cluster of 219 models that score GDTTS of 95 and almost all its residues within 1-Å RMSD of the target (Figures 1A and S2). The other situation, instead, occurs when the top cluster is followed closely by other clusters along the scores, sometimes even swapping rankings. Then, careful inspection of all listed models is required. In our experience, at

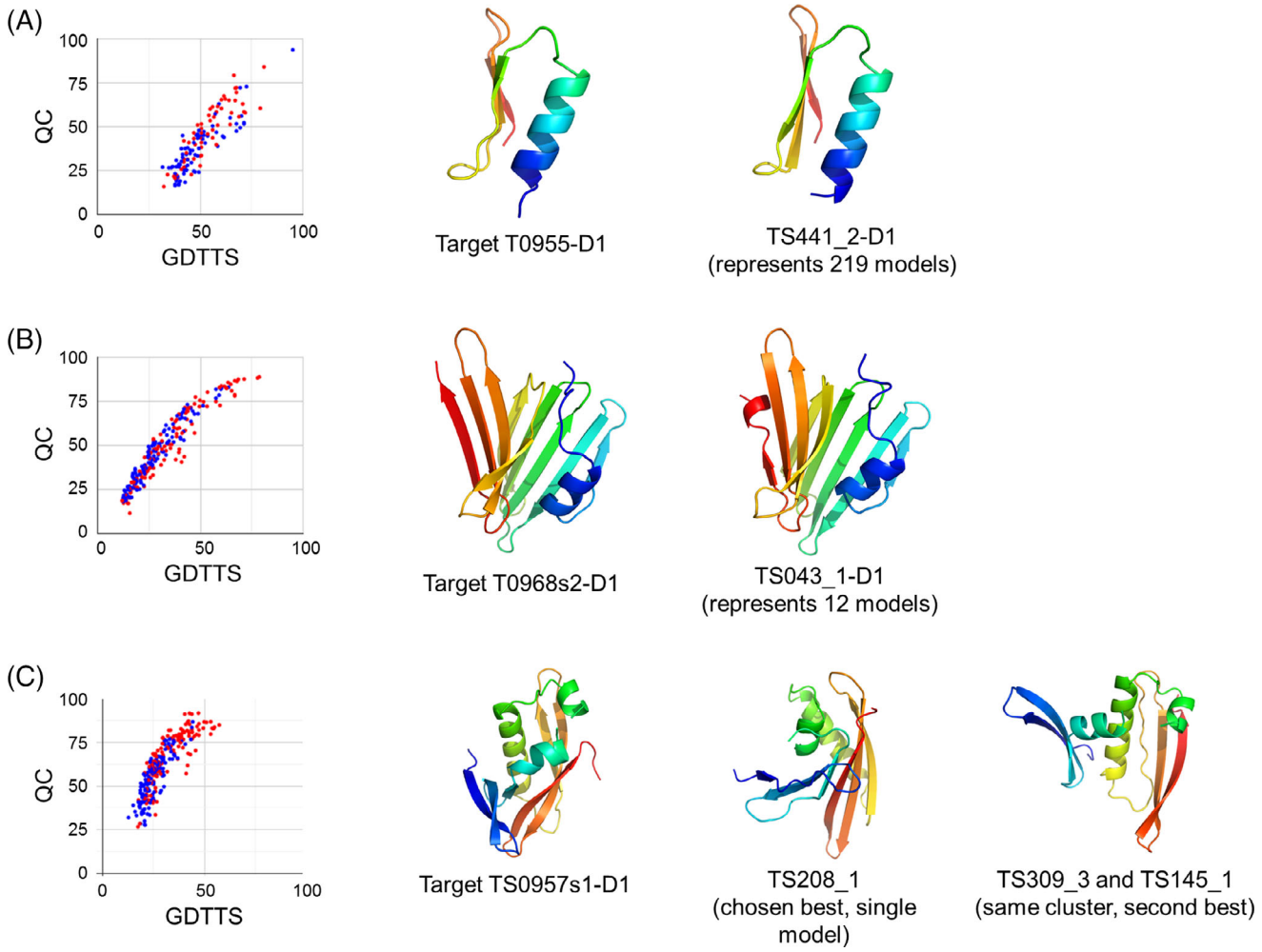

FIGURE 1 Correlation plots for QCS against GDTTS for three of the five evaluation units representative of the kind of situations found upon scores-guided visual evaluation (two additional examples in Figure 2): T0955-D1, T0968s2-D1 and T0957s1-D1. Structures are rainbow-colored from blue at the N-terminus to red at the C-terminus. In the scatter plots, blue dots are clusters that contain at least one server prediction, while red dots are clusters that contain no server predictions. Plots of GDTTS against other scores are shown in Figures S2–S4. GDTTS, Global Distance Test Total Score. All plots can be inspected in detail at http://predictioncenter.org/casp13/casp13-topology-assessment/

GDTTS values above 60-70 one of the top two or three clusters by GDTTS often turns out to be also best upon visual comparison. One example is T0968s2-D1, a β-sandwich with an N-terminal helix for which a large number of models reach high values in all scores, almost 80 in GDTTS and 90 in QCS. The main difference among these several top clusters is the conformation of the C-terminal end (a short α-helix instead of closing up the β-sheet) and the slightly shifted location of the N-terminal helix. Based on this observation, we assign the cluster ranked top by GDTTS (12 models, also top by the other metrics) as the best, whose best model is within 2.33 Å of the target over the full sequence (Figures 1B and S3).

In several cases, most often when the highest scored models reach GDTTS of no more than 50-60, the scores are less correlated. As a result, more clusters are short-listed for evaluation. An extreme case of this situation is when different scores propose totally different top clusters, as illustrated by TS0957s1-D1 where the top nine models by GDTTS rank 6th or worse by QCS and the top five models by QCS rank 13th or worse by GDTTS. Meanwhile, the other scores rather support one or the other in this case, so the short list of clusters for visual inspection does not become too large. For this target, which is a discontinuous FM EU spanning residues 2-37 and 92-163, we readily discarded seven models that do not account for the contacts between these two sequence segments and then assigned TS208 (KIAS-Gdansk)'s model 1 among the rest (single cluster, fifth by Handedness and by CoDM, and eighth by QCS) as the best, as it accounts for the relative position and orientation of the two discontinuous segments better than the other models (Figures 1C and S4). Notice how this model regarded by us as best does not rank top by any score.

Finally, in other cases of low GDTTS we observe that DFM, CoDM, and Handedness support neither GDTTS nor QCS (TM always correlates highly with GDTTS). This lack of correlation results in longer short lists. For example, in T0981-D2 each score proposes three to four different clusters in addition to the five clusters proposed by GDTTS. Thus, a total of 19 clusters deserve inspection (Figure 2A,

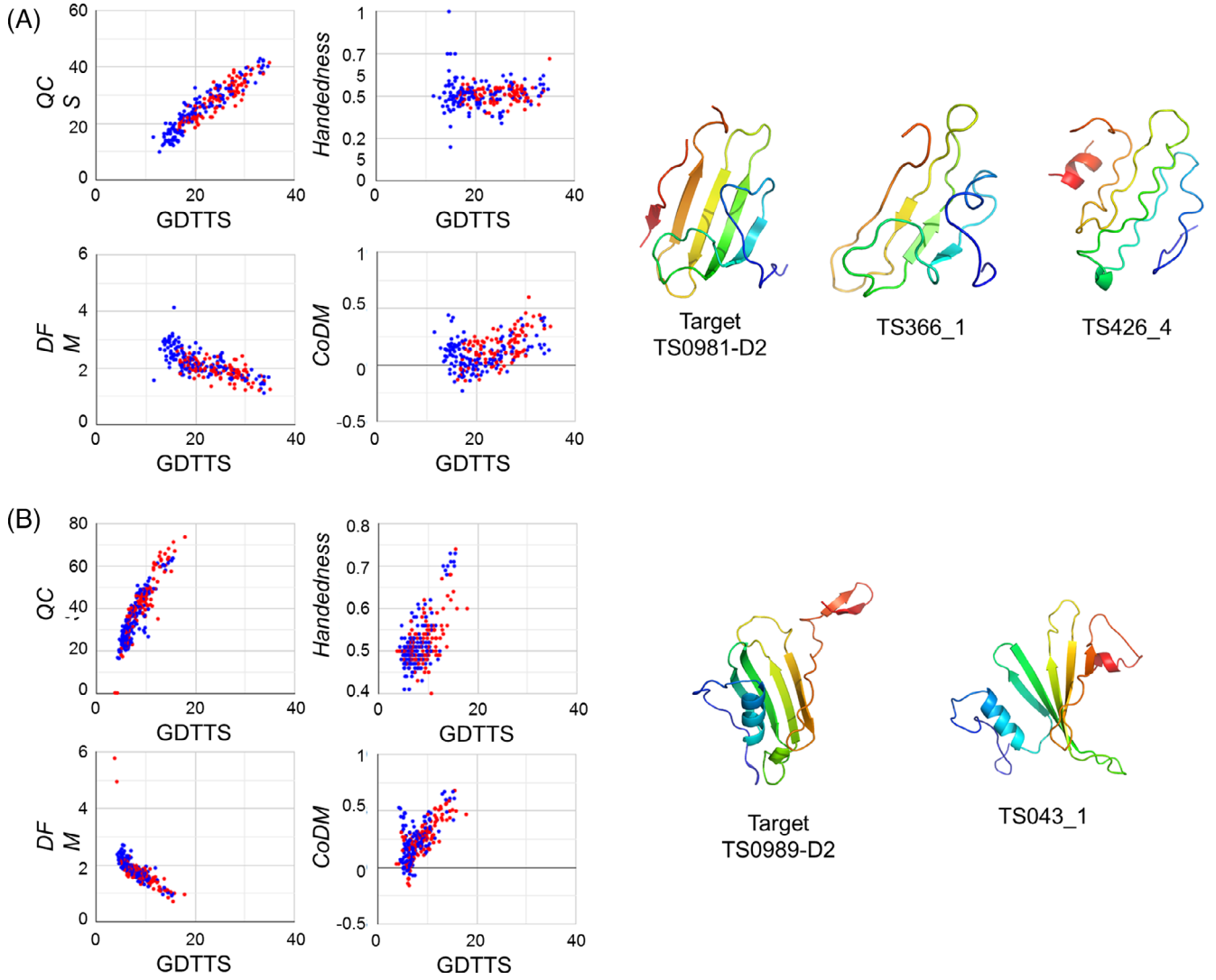

FIGURE 2 Correlation plots for QCS, Handedness, CoDM, and DFM against GDTTS for the two hardest evaluation units of CASP13, also representative of two of the five main kind of situations found upon scores-guided visual evaluation (three other cases in Figure 1). Structures are rainbow-colored from blue at the N-terminus to red at the C-terminus. CoDM, correlation of distance matrices; DFM, deformation; GDTTS, Global Distance Test Total Score. All plots can be inspected in detail at http://predictioncenter.org/casp13/casp13-topology-assessment/

analyzed below). The task becomes even more complex when the highest GDTTS barely reaches 35-45 (below which models are too different from the target and there is nothing positive to rescue) making it very hard to assign best models. This can reach the extreme situation of assigning a tie or just no winner, where it is worth discussing what is good and wrong in each possible best model. Two examples of this, T0981-D2 just mentioned and T0989-D2, constitute what we judged as the two hardest EUs in CASP13, as described next.

3.2 | The hardest FM cases in CASP13

Probably for the first time in CASP, there is at least one model that roughly captures global topology for all EUs. However, we still observe some EUs that remain very challenging.

As anticipated above, the two hardest EUs in CASP13 were probably T0981-D2 and T0989-D2 (Figure 2A,B). T0981-D2 is an FM EU that consists of a six-strand β-sheet twisted on one side, with three ~7-9 residue long loops inserted in-between strands that interact with one face of the sheet; the C-terminal strand of the sheet comes from a separate, downstream segment of sequence, that is, the EU is discontinuous. Model scores barely reach 35 by GDTTS and 43 by QCS, with little overlap in the top clusters by each metric as explained above. Among the top clusters pooled from all scores, we can immediately discard some that do not capture the sequence discontinuity, as well as some fully extended models that artificially rank well by Handedness and CoDM. These short-listed clusters ranked 3rd, 9th, 13th and 19th by GDTTS in the web app for visual analysis, all capturing the C-terminal segment but in an α-helical conformation. Two of them further change strands of the β-sheet by helices, while TS426 (AP_1)'s models 4 and 5 (single cluster ranked 19th by GDTTS) capture the position of the sequence elements best, but reproduce very poorly the β-conformation of the β-strands. On the other hand, the cluster

ranked first by GDTTS (four models from group TS366, Venclovas) does not include the strand coming from the discontinuous C-terminal segment, but captures the β-sheet structure the best. Overall we designate this cluster as the best one if we overlook the discontinuity, or the cluster with TS426's models 4 and 5 as the one that captures the sequence discontinuity the best.

T0989-D2 consists of a five-strand β-sheet with short connecting helices and loops folded on one face of the β-sheet, and a C-terminal β-hairpin that sticks out of the core sheet. GDTTS and TM scores point at a cluster that stands out from the rest, but with a GDTTS of only 45. This cluster is not well supported by the other scores, except QCS where it ranks top at around 75 but followed closely by many other models. Upon visual inspection of the clusters pooled by all scoring metrics, it is clear that the two clusters ranked top by GDTTS do not capture the C-terminal β-hairpin extended away from the core, rather folding it on top of the core β-sheet. Some clusters ranked top by other metrics somewhat capture the C-terminal β-hairpin, at the expense of less defined β-sheets and N-terminal loop. Overall, TS043 (A7D)'s model 1 (single model, third by GDTTS) seems a reasonable compromise between accuracy of the C-terminal hairpin and of the rest of the fold, but we preferred not to assign any top model for this EU.

3.3 | Near atomistic backbone resolution is reached for small targets by several groups

With an average GDTTS of around 62 ± 12 across top GDTTS models for all target EUs (higher than all previous CASPs, see Section 3.6) there are three EUs with model(s) of GDTTS >80 and nine with model(s) of GDTTS >70 in CASP13, against none and only four in the three previous CASPs together. Considering the definition of GDTTS, such models should either have substantial portions between 1 and 2 Å Cα-RMSD from the target accompanied by some less accurate but still quite good regions, or be within around 2 Å RMSD along their full sequences. We then analyzed all the models that we designated best through scores-guided visual inspection in terms of the RMSD and coverage of the best possible fits (information available in the Extended Information section of the web app).

We found 12 best models that reach Cα RMSD <3 Å for >85% of their sequences, of which four are within 1-2 Å RMSD over their full sequences. This is remarkable considering that typical Cα RMSD for folded proteins in submicrosecond molecular dynamics is around 2-3 Å. These 12 EUs and their best models are shown in Figure 3 together with listing of essential data. We note that several of these excellent models were contributed by multiple groups, not just by one or few specialists.

What are the shared features of these target EUs for which models reach near atomistic detail of the backbone? These 12 EUs are small domains ranging in size from 41 to 154 residues; half classified FM and half FM/TBM; characterized by low to null HHpred score but mid to high LGA score (HHpred measuring sequence similarity to the best sequence-level match to the PDB, and LGA the structural similarity of the best available template regardless of the sequence similarity, see domain definition and classification paper by Kinch et al in this issue). Namely, LGA ranges between roughly 50 and 95, meaning that, if found, there are good (in some cases very good) templates in the PDB. Despite these templates being hard to find at the sequence level, their structural information could be contained in the training sets used for informing machine learning methods for contact prediction, which these methods are then able to recall. Besides, HHblits results in nonvoid alignments for all these EUs, which enables prediction of contact maps as witnessed in CASP12 and also from CASP13 abstracts. Still, some remarkable cases like T0955-D1 or T1008-D1 possess rather shallow alignments (although predictors could in principle obtain deeper alignments by using nonpublic sequence databases, eg, from metagenomics projects, this is somewhat unlikely for cases where large numbers of predictors submitted similarly good models). Possibly further helping to produce good models for these 12 EUs, all but two of them (T0990-D1 and T1021s3-D2) correspond to full targets (leaving out short termini trimmed out upon EU definition), which precludes any problems arising from domain definition, discontinuities, and so forth.

3.4 | Cases of good predictions for full targets that EU definition methods suggested splitting

Probably a consequence of the improvements in structure prediction as documented in this article, this CASP features several cases of good predictions for full targets that normal EU-definition methods suggested splitting. This is beneficial to CASP and the modeling community as, of course, the ultimate goal of modeling is to approach full sequences.

One interesting example is T0953s2, a ~250-residue protein that the EU definition process suggested splitting into three EUs. However, we observe that some full models reproduce the overall topology and even some details quite well, especially for EUs D2 and D3. SeeFigure 4A for model 3 by group Destini (224) which is ranked top by GDTTS or model 4 by group Jones-UCL (117) which is ranked top by TM. The latter in particular reaches a Cα RMSD of 2.5 Å over 61% of the full sequence, and its D2-D3 can be superimposed fully within 4.3 Å Cα RMSD.

Examples not only entail exclusively FM and FM/TBM EUs. Target T0957s1 was split in an FM EU (D1) and a TBM-Hard EU (D2, not part of this evaluation track). As we explained above, T0957s1-D1 is a discontinuous target, somewhat difficult to evaluate because different scores propose totally different top clusters. By considering the full target there is better agreement among the scores resulting in fewer clusters for evaluation, and the evaluation is easier. Despite three clusters standing out from the rest with a GDTTS of almost 50 (containing four of group A7D's five models), they do not capture the location of the N-terminus as well as group Jones-UCL's model 4 which ranks fifth by GDTTS (although this model does not capture very well the secondary structures) (Figure 4B).

A more extreme example not entailing exclusively FM and FM/TBM EUs is that of targets that contain only TBM EUs, like T1014 which consists of two globular TBM-easy EUs docked to each other and connected through a long linker. Naturally, most models are

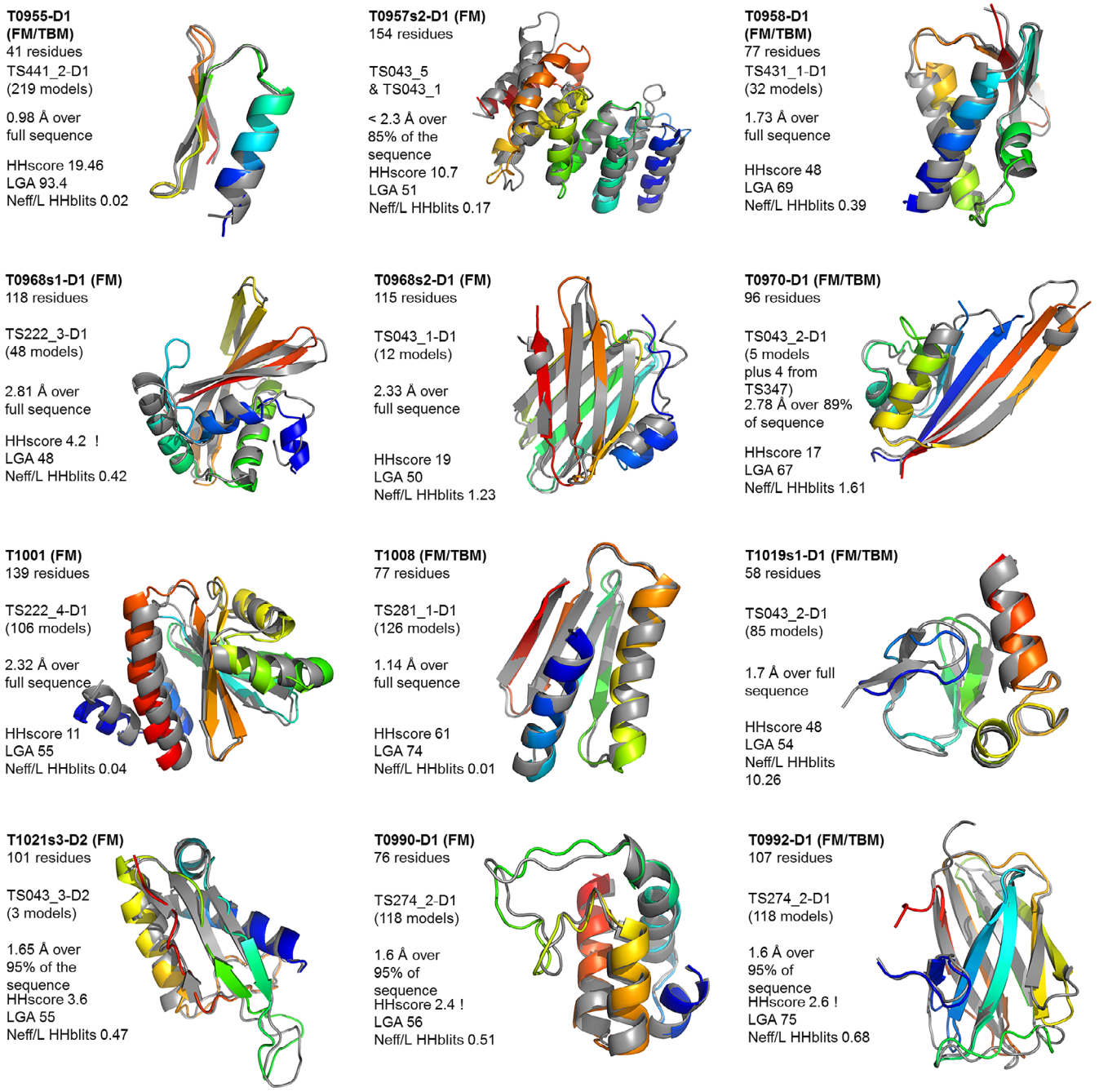

FIGURE 3 The 12 evaluation units for which best models reproduce the Cα trace at nearly atomistic detail. All targets are rainbow-colored and models shown in gray. Essential information relevant to the text is provided

very good for the individual domains; however, the relative orientation, docked pose, and interactions between the two domains are quite bad, such that all but five models barely reach a GDTTS of 50. The five good models (one by group BAKER (086) and four by chuo-u (047)) stand as a single cluster of GDTTS 58.7; they are not superb, as they capture the relative orientation of the two domains rather well but with a significant rotation that implies incorrect interactions (Figure 4C).

In these and other cases explored, Grishin plots[8] do suggest splitting (although the decision is not based only on these plots but taken together with information from domain parsers, coverage of available templates, and inspection of the target structure itself), but we observed in some cases trace evidence suggesting no splitting. Recalling that Grishin plots are built from the weighted sum of GDTTS scores for the individual EUs vs the GDTTS of the unsplit target taking the top 20 server models only, we expected that since now models are getting so good, especially from human groups, it could be more informative to look at Grishin plots built from all the submitted models. As a first case we will look at target T1000, which was split into two EUs: one for residues 10-92 that is not evaluated because

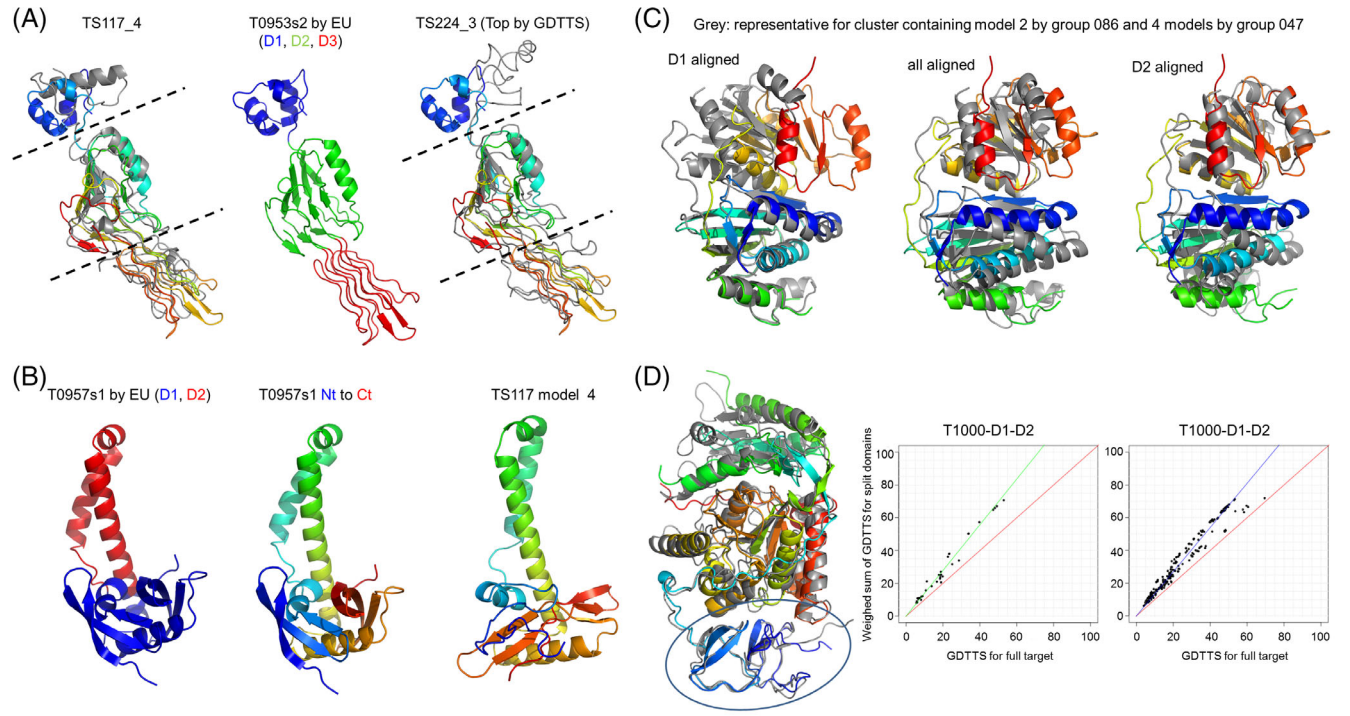

FIGURE 4 Examples of analysis of full targets for T0953s2 (A), T0957s1 (B), T1014 (C), and T1000 (D). When superimposed, targets are colored from blue at the N-terminus to red at the C-terminus, and models are colored gray. In D, the oval highlights the domain for which an X-ray structure is available in the PDB, and the plots are Grishin plots for T1000's D1 and D2 considering only server predictions (left) or all models (right). PDB, Protein Data Bank

there is a PDB structure for a protein with essentially the same sequence and structure, and the other FM EU for residues 93-523 that spans the rest of the target. We note that predictors submit predictions for full sequences, though, and that similarly to the case of T1014 shown above, models should reproduce how the non-evaluated, N-terminal domain integrates into the rest of the structure. As described in the Annex 1 of Data S1, we designated two clusters as best for the FM EU alone; the same two clusters recapitulate the full structure similarly well, possibly slightly better by the cluster ranked top by all metrics (model 1 by A7D) which has a GDTTS of 70 meaning that it is very good, in fact it has 89% of its residues within 2 Å of Cα RMSD to the full target (Figure 4D). The server-only Grishin plot clearly suggests separation; however, the all-group Grishin plot features a substantial fraction of models reaching high GDTTS on both axes and on the diagonal, suggesting no splitting.

As a consequence of observing good models for a handful of full targets, some were included as cases of FM-special EUs after discussion with the group doing EU definition and classification. We feel confident to propose that in future CASP editions the focus should be on evaluating full targets, not EUs or domains, in the tertiary structure track. Not only when the all-group Grishin plots support this, but also when they do not, because, for example, targets made up of two TBM units will almost always hit above the diagonal as each domain can be perfectly modeled (see, eg, the Grishin plot for target T0984 which has two TBM-easy EUs in Figure S5).

Splitting into EUs probably made perfect sense in early CASPs to facilitate rescuing some correctly modeled features within overall bad models; but now, with the improvements we have witnessed, we think CASP can move on to the evaluation of what is more important regarding the biological questions, that is, the structures of full targets. However, domains connected by flexible linkers will still require special attention and splitting might be needed, especially when domains seem loosely bound and might rather explore a continuum of docked poses one of which just got favored upon crystallization.

3.5 | Best performing groups, official CASP ranking, and notable highlights

With our scores-guided evaluation completed for all EUs, we can approximate a first ranking by counting how many models each predictor submitted that we judged as best. In doing so, group A7D (043) stands out with 31 expert-judged best models followed by MULTICOM (089) with slightly over half the counts and then by several runners up that follow closely (Figure 5A). Although there is no yet any official publication on A7D, the interested reader is referred to a recent article by AlQuraishi[17] presenting some aspects of A7D's methods as well as other interesting considerations.

A ranking based on the number of models judged as best is, however, not fair because (a) it does not reward very good predictions that would be second or third best (to the extreme that a group submitting all models that are second-best would score 0!) and (b) it does not penalize groups that submitted bad models for EUs for which many groups submitted good models. These and other problems are sorted

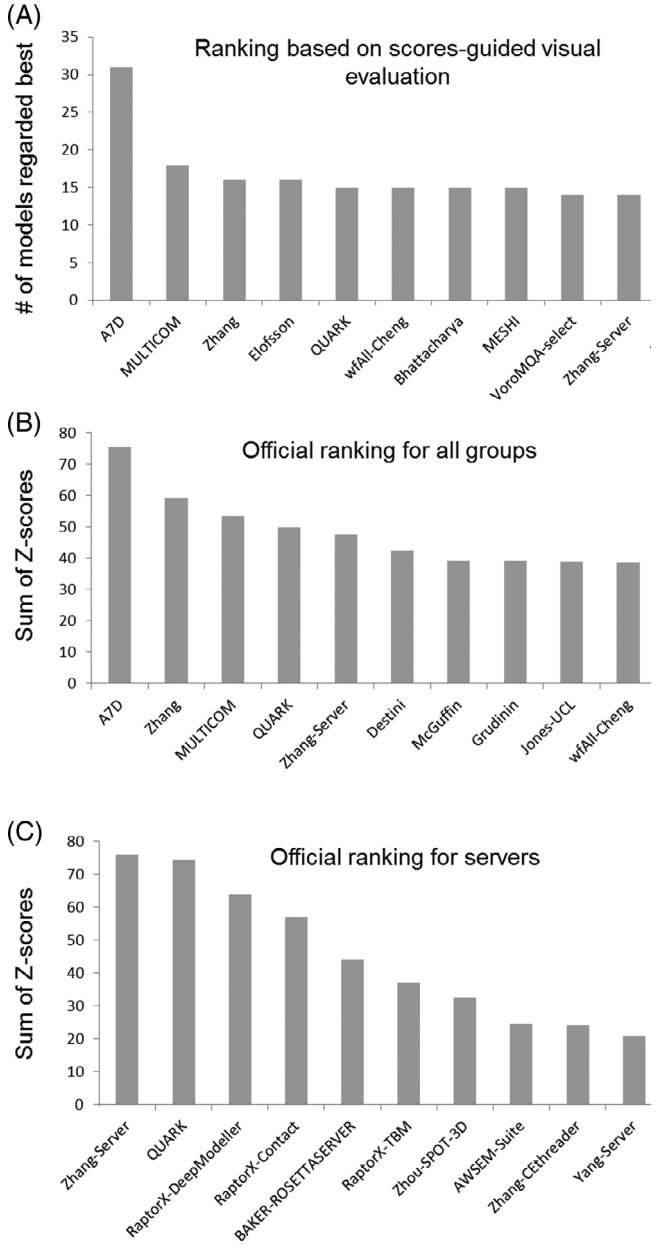

FIGURE 5 Predictors ordered by the number of models regarded as best by our scores-guided visual evaluation (A) and official ranking separated for all groups (B) and servers (C)

out in CASP by ranking groups based on *Z*-scores of the most relevant quality metrics, most often GDTTS alone or sometimes combined with other scores, averaged throughout all EUs. Our experience during scores-guided visual inspection in CASP12 and CASP 13 has been that most often, GDTTS and QCS are the two most informative scores and are very complementary. Although this remains a subjective statement based on our experience, it is backed up more objectively in the analysis of several structure-comparison scores by Olechnovic et al.[18]

We therefore produced the official CASP13 ranking by summing up *Z*-scores combined from GDTTS and QCS (equally weighted) on all models submitted as #1, for TBM/FM, FM, and FM_sp target EUs, and considering sum of *Z*-score > −2. The final official ranking of predictors in the CASP13 track of tertiary structure prediction is shown separately for all groups and servers in Figure 5B,C. We point out that the top five predictors among all groups (A7D, Zhang, MULTICOM, QUARK, and Zhang-server) remain the same and in the same order if the *Z*-scores are computed on GDTTS alone as often done in CASP, or on QCS that is, the other metric we consider of most importance for visual evaluation, or on the widely used TM score.

Another important observation, which we made also in CASP12, is that some of the groups that do not rank at the top still are the ones who deliver best models for some specific EUs. In CASP13, these were ZHOU-SPOT for T0998-D1, their best model being unique and quite better than runners-up; Jones-UCL for T1010-D1, their best model also being unique and better than runners-up; RaptorX-DeepModeller for T0949 although here it was harder to tell among several relatively good models that all missed the same portion (where this group's model achieves the most informative structure, see Annex 1 of Data S1 for details); KIAS-Gdansk for T0957s1-D1; BAKER for T0975-D1; and Venclovas for T0991-D1. We further highlight some best models for full targets contributed by Jones-UCL for T0953s2 and T0957s1, and by BAKER and chuo-u for T1014.

We finally highlight that the groups ranking at the top or being especially good for specific targets rely largely on methods for contact or distance predictions including a substantial contribution from novel machine learning methods, and instead there is very limited contribution from physics-based methods that achieved notable results for some targets in CASP12.

3.6 | Progress in tertiary structure prediction

Our assessment has so far shown that tertiary structure prediction of hard targets went well in CASP13, with the following reasons suggesting a sizable improvement relative to the previous CASP editions: (a) CASP13 features at least one model which roughly captures global topology for all EUs, probably unprecedented in CASP; (b) top models reach backbone RMSD <3 Å for several EUs of size <150 residues, contributed by many groups; and (c) there are even quite good models for full, unsplit targets.

To track the progress in tertiary structure prediction of the hardest targets in a more quantitative way, we have collected the models of highest GDTTS for all FM EUs (or equivalent in early CASPs). Figure 6A shows the median GDTTS of FM EUs per CASP over time, extending the plot we presented in our previous CASP12 evaluation; while Figure 6B shows the average of GDTTS values. Both graphs show a clear improvement in the last three CASPs with CASP13 featuring the best predictions ever. We note that CASP13's FM EUs are of similar difficulty to those of CASP12 as judged from EU sizes and difficulty metrics shown in Table 1 making the improvement from CASP12 to CASP13 more significant. This is also patent when the range of LGA values (which measure structural similarity to templates, green rhomboids in Figure 6) for CASP12 and CASP13 are compared to the maximum GDTTS. Furthermore, such comparison shows large overlap between LGA and highest GDTTS in CASP12, broken in

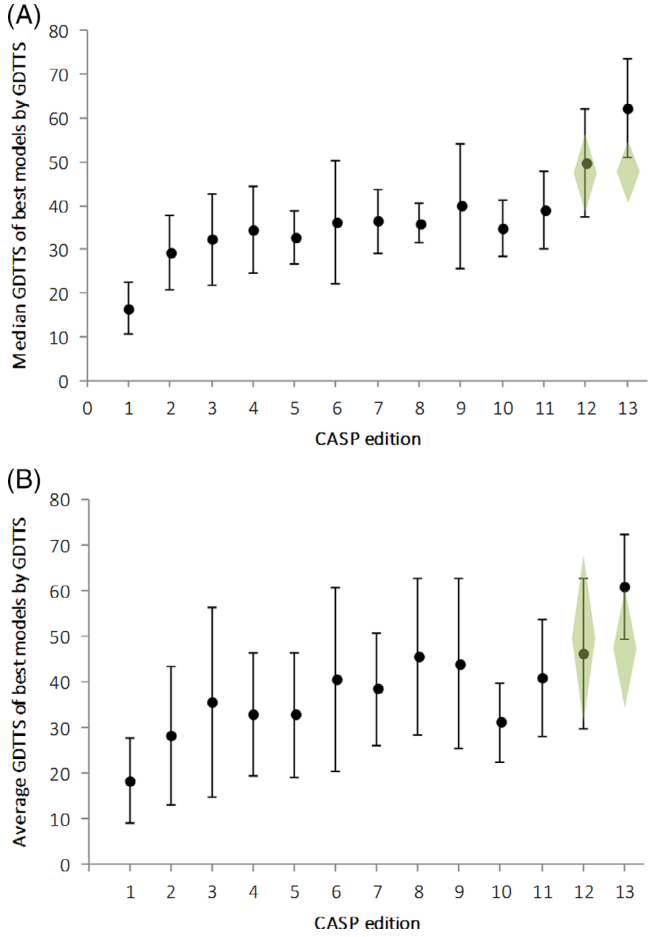

FIGURE 6 Progress in the tertiary structure prediction track of CASP since its first edition, quantified by the median (top) and average (bottom) of the highest GDTTS observed for all FM evaluation units (or equivalent definitions in early editions). The green shades for CASP12 and 13 indicate the median ± median absolute deviation (top) and average ± SD (bottom) of the distributions of LGA scores for the top templates identified by Kinch et al. CASP, Critical Assessment of Structure Prediction; FM, Free Modeling; GDTTS, Global Distance Test Total Score

CASP13 which suggests improvement of the models relative to the structurally closest templates (see more details in Figure 7D).

What are the possible causes for improved structure predictions? In CASP12, the main improvements in modeling of hard targets were brought about by the initial prediction of residue-residue contacts, in which an independent assessment showed had improved substantially from CASP11.[19] According to CASP13 abstracts, such predictions were also important for modeling and, moreover, they have experienced a further substantial improvement (article by the Fiser group in this issue). Naturally, better contact predictions could explain better models; however, we noticed from the CASP13 abstracts that some groups (notably A7D, MULTICOM-related and RaptorX-related groups) went one step further and also predicted distances between pairs of residues, as long as 10-20 Å. Such predictions can provide information on the relative arrangement in 3D of multiple residues, which if accurate should facilitate predictions even further than plain

TABLE 1 Difficulty of CASP13 FM EUs compared to CASP12 FM EUs, as quantified with 6 metrics

Metric	CASP12	CASP13
HHpred	11.7 ± 15.3	10.0 ± 10.0
LGA	50.5 ± 17.9	47.9 ± 13.4
(HHpred + LGA)/2	31.1 ± 11.6	28.9 ± 8.7
Server GDTTS	27.3 ± 10.1	31.2 ± 8.3
Length	154 ± 80	155 ± 82
Log (1 + *N*/*L*)	0.35 ± 0.48	0.19 ± 0.23

Abbreviations: CASP, Critical Assessment of Structure Prediction; EUs, evaluation units; FM, Free Modeling; GDTTS, Global Distance Test Total Score.

contact maps (predictor-contributed papers in this issue should illuminate more on this).

Overall, the progress plots seem to highlight a first "wave" of improved contact predictions and models caused by coevolution methods between CASP10 and CASP12, and then a second "wave" of improved contact (and now also distance) predictions and models fostered by integration of coevolution data, sequence-wise physicochemical features, and even structures through very modern machine learning approaches. Given that these methods rely strongly on the availability of sequence alignments, we next analyze the impact of alignment depths on model quality and compare this to CASP12. Figure 7A,B shows plots of the highest GDTTS reached for each target EU against the log (Neff/*L*), for CASP12 and CASP13, separately for FM and FM/TBM (Neff being the number of sequences retrieved in alignments and *L* is the EU's length in residues). We observe that best models are better for EUs for which deeper alignments are available, just like in CASP12 and indeed with a similar slope in the linear trend. However, the intercept is around 13 GDTTS units higher in CASP13, suggesting that methods are now performing better and working as well with shallower alignments, thus becoming useful for more proteins.

We finally note that while deep enough alignments might help to fold very hard targets, the best models are still those of small EUs; in fact we observe some negative correlation between highest GDTTS and EU size for both FM and FM/TBM EUs (Figure 7C). Notably, some of the very good predictions are for EUs that have relatively high LGA score despite low HHpred (see examples in Figure 3), which might reflect the impact of training sets used to inform machine learning methods for contact prediction, such that these methods can "recall" the contact patterns contributed by those PDB entries to the training sets, even if for completely different sequences. Correspondingly, there is a positive correlation between the highest GDTTS scores and LGA difficulty metric for FM/TBM EUs (Figure 7D); and interestingly there is no such correlation for FM models even though some reach LGA as high as 70-80.

3.7 | New routes for future CASPs

Our progress analysis highlights substantial improvements in the last two CASPs, especially in CASP13, which lead us to propose new

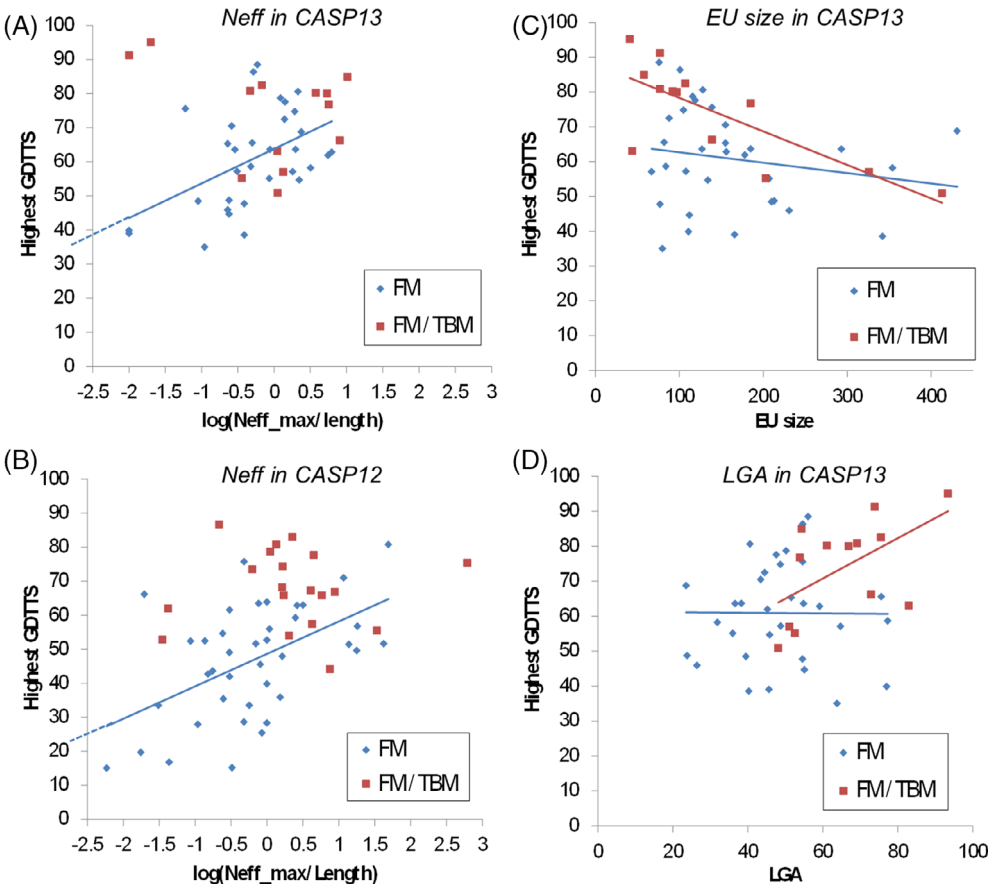

FIGURE 7 Plots correlating highest GDTTS with alignment depths for EUs in CASP13 (A) and in CASP12 (B); and with EU sizes (C, in numbers of residues) and LGA difficulty metric (D) for CASP13 targets CASP, Critical Assessment of Structure Prediction; EUs, evaluation units; GDTTS, Global Distance Test Total Score

routes for evaluation in the tertiary structure prediction track. First, as advanced in previous sections, the time to drop splitting into EUs has probably arrived, so that future assessors shall assess full targets as the main entities for evaluation. This will allow evaluation of more biologically relevant models of tertiary structures, while the other relevant entity, the quaternary structure, is already evaluated in another track of CASP. Importantly, everything the CASP organizers and assessors have learned about EU definition and classification throughout the history of CASP will still be useful, for example, to split extremely complex cases and thus allow assessors to decide what was correctly modeled and why, to remove flexible linkers or to dissect loosely bound domains that may have been artifactually stabilized by crystal packing. Regarding EU classification by difficulty, including alignment depths or the quality of some publicly available standard of contact predictions could be added on top of HHpred and LGA metrics currently used.

Second, the state of the art of tertiary structure predictions by CASP13 further suggests that we could now move on to harsher analyses focused on finer details. In particular, getting backbones right within ~2 Å Cα RMSD calls for an analysis of the local details such as packing and side-chain conformations. Much of this more detailed assessment could probably be based on the most established protocols for the assessment of models of TBM EUs.

Third, another important point is assessing quality estimate predictions, which models are supposed to contain at the residue level coded in their B-factor column. For example, if two equally good models differ in their estimation of the errors, one correctly estimating an incorrectly modeled region but the other one estimating it as correct, then the former should be scored better. Our web app allows inspecting this by navigating the models within a cluster colored by their B-factor columns, but future work would require an investigation on how to properly balance quality estimation and actual similarity of the model to the target. We next present two examples that were easy to spot by navigating models of the cluster with our web app. The first case (Figure 8A) is the cluster of models designated the best for T0955, which contains 219 models among which (i) some (including that of absolute highest GDTTS) have no quality estimate, (ii) some estimate that the short loops might deviate from the target, and (iii) others estimate that not only loops but also the most exposed residues of the β-sheet and α-helix, or even the whole α-helix, might be off. Here, we would stand in favor of case (ii) as they correctly predict the fold and are certain about this. The second easy-to-spot example is that of T1001-D1's best cluster (Figure 8B), whose model of highest absolute GDTTS estimates two whole β-strands to be off (despite they are correct) and a central helix to be perfect, while the runner-up within the cluster is as good but estimates the two β-strands to be correct and warns that the central helix (which is as wrong as in the other model) might be off especially on its C-terminal side, as indeed observed. This model (model 1 by group 117, Jones-UCL) actually seems the best regarding quality estimate within this

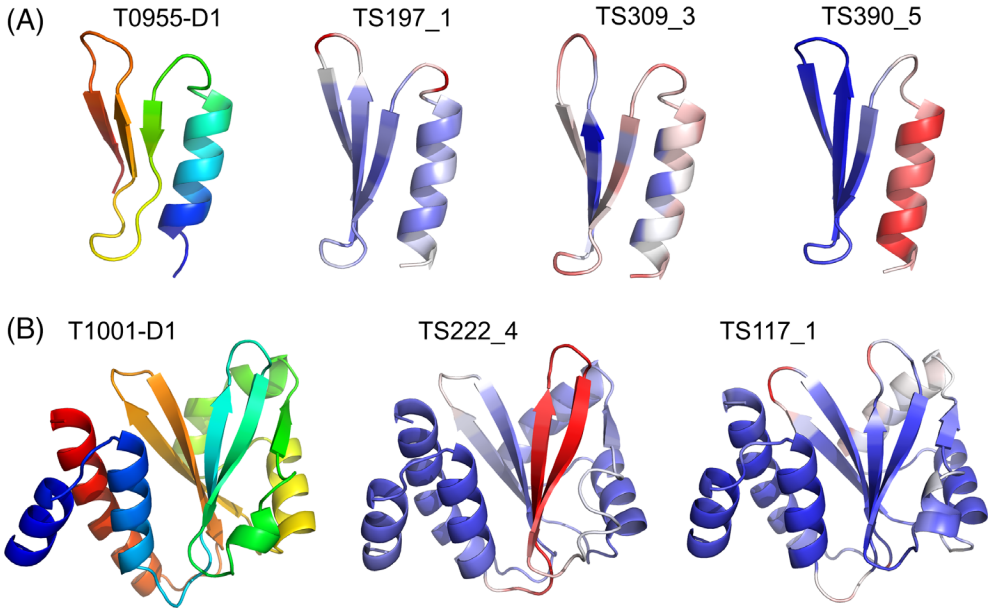

FIGURE 8 Inspection of quality estimates in models of the cluster selected as best for T0955-D1 (A) and that for T1001-D1 (B). Target evaluation units are rainbow-colored from blue to red, and models are colored according to their B-factor columns (ie, residue-wise quality estimates) from blue for 0 to white to red for the maximum value

cluster, so if this was to be considered for evaluation we would probably designate it as best, alone.

As a further extension of quality estimation assessments, because many experiments performed in the wet lab to infer clues on a protein's structure often involve the relative positions of two residues (such as engineering disulfide bonds, paramagnetic probes, or fluorophores for FRET experiments), in the future CASP should consider asking predictors for residue-residue quality estimates, to be considered upon evaluation.

We think CASP is ready to undertake some if not all of the above propositions already from next edition. We expect this will certainly add great value to the tertiary structure prediction track and bring it closer to its actual utility in biology. Other more advanced predictions often discussed in CASP, such as evaluating conformational landscapes rather than static structures, are likely still farfetched. The later possibility is, however, exciting especially as cryo-electron microscopy evolves to capture multiple conformational states[20] and as the technique contributes more CASP targets.

3.8 | Still open problems in FM

Following from a similar section in our previous CASP12 assessment,[3] the CASP13 targets highlight some still open problems in tertiary structure prediction:

- For targets longer than 150 residues, models tend to be better than in CASP12, but they are still not as good as models of smaller targets.
- Relative to CASP12, groups seem to have improved in identifying templates that have medium to high structural similarity to the target despite low sequence similarity, as exemplified in several cases above especially when presenting the near-atomistic models.
- As in CASP12, there are some targets with very shallow alignments; and although the predictions seem to have improved and require less sequence to achieve a similar quality, the dependence on alignment depth is still clear.
- Segments of undefined secondary structure as in the several cases from Figure 4 are hard to match to the target. While part of this could be due to true dynamics, several deviations seem too large to be so. Maybe this and other kinds of deviations could be better assessed considering quality estimates in the future.
- Domain-domain interactions within targets are not very accurately modeled, even if the domains themselves are nearly perfect, as in the case of T1014.

4 | CONCLUSIONS

In CASP13, we have witnessed yet another step forward in the modeling of difficult targets, from at least some model capturing the coarse structure of the hardest targets to models with minimal backbone deviations. Also good models were submitted for full targets that domain-definition methods suggested splitting.

Predictors globally ranking top for the most difficult targets were the A7D, Zhang, and MULTICOM groups. Compared to CASP12, this time physics-only-based methods were less successful in contributing best models for specific targets. Machine learning-based methods for prediction of residue-residue contacts, and now distances, currently outperform the physics-only methods by far.

Like in CASP11 and CASP12, FM predictions seem to benefit largely from contact prediction methods through machine learning methods, but now working better at similar alignment depths, or similarly well on shallower alignments. CASP13 further sees the introduction of machine learning for residue-residue distance predictions, which being more informative than mere contact predictions may

have also helped to improve models. As both contact and distance predictions depend today heavily on alignments, alignment depths might be considered as additional metrics of target difficulty in future CASPs.

Compared to previous CASPs, the 150 residue limit seems to not have been broken, and there is still a clear dependence of model quality on protein size, structural similarity to PDB templates, and alignment depth for FM targets. We therefore see large space for progress, despite the very evident and encouraging improvements. Tighter coupling between the tracks for tertiary and quaternary structure predictions, or even with CAPRI, might be required, as large proteins often contain multiple domains docked to each other much like the protomers of multimeric assemblies but joined by a covalent flexible linker (as in targets T1000 or T1014, Figure 4C,D).

Our assessment allows us to propose taking this track of CASP to a new level, already from CASP14, by dropping (most) EU splitting and focusing on full targets, evaluating packing and sidechains for models with excellent backbone traces, and stressing attention on assessing quality estimates, all bringing the CASP experiment closer to the reality of the actual structure predictions as needed in biology.

ACKNOWLEDGEMENTS

We thank the CASP13 organizers for the invitation to participate once more as assessors for the topology prediction track, to the other assessors of this and previous CASPs; and especially to B. Monastyrskyy and A. Kryshtafovych from the Prediction Center (University of California) and Lisa Kinch (HHMI and University of Texas) for much appreciated help, and to the groups that contributed structures, experimental data, and predictions for CASP13. M.D.P. thanks for the support of the EPFL and the Swiss National Science Foundation.

CONFLICT OF INTERESTS

The authors declare no potential conflict of interest.

ORCID


Luciano A. Abriata https://orcid.org/0000-0003-3087-8677

REFERENCES

1. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—progress and new directions in Round XI. *Proteins*. 2016;84:4-14.
2. Kinch LN, Kryshtafovych A, Monastyrskyy B, Grishin NV. CASP13 target classification into tertiary structure prediction categories. *Proteins*. 2019. https://doi.org/10.1002/prot.25775.
3. Abriata LA, Tamo G, Monastyrskyy B, Kryshtafovych A, Dal Peraro M. Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins Struct Funct Bioinforma*. 2018;86(Suppl 1):97-112.
4. Tai C-H, Bai H, Taylor TJ, Lee B. Assessment of template-free modeling in CASP10 and ROLL. *Proteins*. 2014;82(Suppl 2):57-83.
5. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. *Proteins*. 1999;3:22-29.
6. Abriata LA. Web apps come of age for molecular sciences. *Informatics*. 2017;4:28.
7. Abriata LA, Rodrigues JPGLM, Salathé M, Patiny L. Augmenting research, education and outreach with client-side web programming. *Trends Biotechnol*. 2017;36(5):473-476.
8. Kinch L, Yong Shi S, Cong Q, Cheng H, Liao Y, Grishin NV. CASP9 assessment of free modeling target predictions. *Proteins*. 2011;79 (Suppl 10):59-73.
9. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins*. 2004;57:702-710.
10. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*. 2010;26:889-895.
11. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP11 statistics and the prediction center evaluation system. *Proteins*. 2016;84(Suppl 1): 15-19.
12. Hanson RM, Prilusky J, Renjian Z, Nakane T, Sussman JL. JSmol and the next-generation web-based representation of 3D molecular structure as applied to Proteopedia. *Israel J Chem*. 2013;53:207-216.
13. 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;29:2722-2728.
14. Antczak PLM, Ratajczak T, Blazewicz J, Lukasiak P. SphereGrinder—reference structure-based tool for quality assessment of protein structural models. *IEEE Int Conf Bioinforma Biomed BIBM*. 2015;1: 665-668.
15. Chen VB, Arendall WB, Headd JJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr*. 2010;66:12-21.
16. Abriata LA, Kinch LN, Tamò GE, Monastyrskyy B, Kryshtafovych A, Dal Peraro M. Definition and classification of evaluation units for tertiary structure prediction in CASP12 facilitated through semi-automated metrics. *Proteins*. 2017;86:16-26.
17. AlQuraishi M. AlphaFold at CASP13. *Bioinformatics*. 2019. https://doi.org/10.1093/bioinformatics/btz422.
18. Olechnovic K, Monastyrskyy B, Kryshtafovych A, Venclovas Č. Comparative analysis of methods for evaluation of protein models against native structures. *Bioinformatics*. 2018;35(6):937-944.
19. Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin AMJJ. Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. *Proteins*. 2017;86(Suppl 1):51-66.
20. Frank J. New opportunities created by single-particle Cryo-EM: the mapping of conformational space. *Biochemistry*. 2018;57:888.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Abriata LA, Tamò GE, Dal Peraro M. A further leap of improvement in tertiary structure prediction in CASP13 prompts new routes for future assessments. *Proteins*. 2019;1–13. https://doi.org/10.1002/prot.25787