Received: 11 June 2021 | Revised: 6 September 2021 | Accepted: 21 September 2021

DOI: 10.1002/prot.26248

RESEARCH ARTICLE

PROTEINS WILEY

Assessing the accuracy of contact and distance predictions in CASP14

Victoria Ruiz-Serra[1] | Camila Pontes[1] | Edoardo Milanetti[2,3] | Andriy Kryshtafovych[4] | Rosalba Lepore[1] | Alfonso Valencia[1,5]

[1]Barcelona Supercomputing Center (BSC), Barcelona, Spain

[2]Department of Physics, Sapienza Università di Roma, Rome, Italy

[3]Center for Life Nano- & Neuro-Science, Fondazione Istituto Italiano di Tecnologia (IIT), Rome, Italy

[4]Genome Center, University of California, Davis, California, USA

[5]ICREA, Pg. Lluís Companys, Barcelona, Spain

Correspondence
Rosalba Lepore, Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain.
Email: alba.lepore@bsc.es

Funding information
National Institute of General Medical Sciences, Grant/Award Number: GM100482

Abstract

We present the results of the assessment of the intramolecular residue–residue contact and distance predictions from groups participating in the 14th round of the CASP experiment. The performance of contact prediction methods was evaluated with the measures used in previous CASPs, while distance predictions were assessed based on a new protocol, which considers individual distance pairs as well as the whole predicted distance matrix, using a graph-based framework. The results of the evaluation indicate that predictions by the tFold framework, TripletRes and DeepPotential were the most accurate in both categories. With regards to progress in method performance, the results of the assessment in contact prediction did not reveal any discernible difference when compared to CASP13. Arguably, this could be due to CASP14 FM targets being more challenging than ever before.

KEYWORDS
CASP14, community-wide experiment, benchmarkin, prediction of residue–residue contact and distance, numerical evaluation measures

1 | INTRODUCTION

Contact prediction has been an active area of research since 1994[1,2] and an integral part of CASP since its early days.[3] Much of the research in this area has been inspired by the hypothesis of coevolution, suggesting that compensatory mutations between pairs of amino acids in the MSA of a protein family can be used as a marker of their physical proximity in the 3D structure.[1,4] A number of studies back in the 1990s illustrated the use of contact maps as constraints for protein structure prediction.[5–7] During these two decades we have seen continuous progress in the quality of the predictions by a combination of improvements in the basic sequence correlation algorithms, and better alignments.[3,8–15]

In CASP11, the average precision on L5 long-range contacts in free-modeling targets (FM) reached 27%, mainly driven by the ability of new methods to disentangle direct and indirect coevolutionary signals, that is, the direct coupling analysis (DCA).[16,17] The precision nearly doubled in CASP12,[18,19] thanks to the integration of deep neural networks and the increased availability of sequence data from metagenomics sequencing. In CASP13, another leap in performance raised the limit of contact prediction accuracy to 70%. This was the result of a community-wide adoption of fully deep residual neural networks, able to capture higher-order residue correlations from the global network of contact restraints, and more specifically, of inter-residue distances.[20–22]

While predictions have been usually assessed by measuring the accuracy of pairwise contacts, recent work, including the results of

Abbreviations: CASP, critical assessment of protein structure prediction; DCA, direct coupling analysis; DG, diameter of gyration; DNN, deep neural network; ES, entropy score; FM, free modeling; MBN, mean bin neighbor; MDD, mean distance difference; MSA, multiple sequence alignment; TBM, template-based modeling.

Rosalba Lepore and Alfonso Valencia should be considered joint senior author.

the CASP13 experiment, showed that 3D structure prediction methods can benefit from predictions of inter-residue distances as constraints in the folding algorithms. In particular, the finer-grained information contained in the distance matrix provides more physical constraints and a richer training signal than a contact matrix, which in turn may lead to more accurate predictions of the 3D structure as discussed in papers.[21,23] CASP14 added this new category to the assessment, which in turn required the development of the new assessment methodology presented here. While the employed assessment metrics and procedures were different in the two cases, the results indicate that predictions submitted by Tencent AI Lab and Zhang lab were the most accurate in both categories.

2 | MATERIALS AND METHODS

2.1 | Overview of targets and participating groups

In CASP14, the contact and distance prediction category included a total of 38 targets, 23 belonging to the free modeling (FM) category and 15 to the overlap of free modeling and template-based modeling category (FM/TBM).[36] The size of targets ranged between 72 and 464 amino acids. A total of 60 groups submitted contact predictions, including 51 groups who predicted at least 37 targets and 47 groups who predicted at least 20 targets. Thirty-nine out of the 60 groups also provided distance predictions, with all groups except two predicting at least 37 targets. Detailed information on groups and predictions is provided in Table S1.

2.2 | Prediction format

Definitions, formats, and procedures in the CASP14 contact prediction category did not differ from previous experiments and therefore we provide here only the basic information, encouraging readers to refer to previous CASP assessment papers[15,16,20,24] for more detailed explanations.

Contact predictions were provided in a 3-column format: *i*, *j* and *p*0, where *i* and *j* are amino acid indices, and *p*0 is the contact probability score [0;1]. Residues *i* and *j* are defined to be in contact if d_{ij} < 8.0 Å, where d_{ij} is the distance between their C_β atoms (C_α in case of Glycine). The assessment was performed on the top *L*/*N* contacts (*L* – protein length in residues, *N* = {1,2,5}) according to the contact confidence score *p*0.

Distance predictions were provided in a 13-column format, containing *i*, *j*, and *p*0 (see above) and, additionally, the probabilities *pN* [0;1], reflecting the confidence of inter-residue distance falling within the bin *N*. Distance bins were defined in the increment of 2 Å, with the exception of bin *N* = 1 and bin *N* = 10, with the following boundaries: bin_1: d_{ij} ≤ 4 Å, bin_2: 4 < d_{ij} ≤ 6 Å, bin_3: 6 < d_{ij} ≤ 8 Å, …, bin_{10}: >20 Å. Different from what was done in the contact assessment, distance predictions were assessed based on the entire distance map.

Previous to the evaluation of both contacts and distance predictions, submissions were trimmed to domains and amino acid pairs were excluded if their separation along the sequence was smaller than 6 amino acids. Any non-listed amino acid pair was assumed to be not in contact and assigned *p*0 = *0* for the contact evaluation, or belonging to the last distance bin and assigned *p*10 = 1 for the distance evaluation.

2.3 | Assessment metrics

The assessment of *contact predictions* was performed based on precision, recall and *F*1 score metrics, computed as follows:

$$\text{precision} = \text{TP}/(\text{TP}+\text{FP})$$

$$\text{recall} = \text{TP}/(\text{TP}+\text{FN})$$

$$F1 = \frac{2(\text{precision})(\text{recall})}{\text{precision}+\text{recall}}$$

where TP is the number of true positives (*p*0 > 0 and d_{ij} < 8 Å), FP is the number of false positives (*p*0 > 0 and d_{ij} ≥ 8 Å), and FN is the number of false negatives (*p*0 = 0 and d_{ij} < 8 Å). Contacts were grouped into short, medium, and long-range categories according to their sequence separation, defined as pairs of residues separated by 6–11 residues, 12–23 residues, and 24 residues or more, respectively.

Another metric traditionally employed to evaluate contact predictions is the entropy score (ES), computed as the relative drop of entropy in the protein structure as geometric constraints are imposed according to correctly predicted contacts, as follows[20]:

$$\text{ES} = 100\frac{E(0)-E(C)}{E(0)}$$

where *E*(*C*) and *E*(0) are the entropies of the protein with and without structural constraints, respectively. These entropies are calculated as the average value of the Shannon entropy for residue-residue distances, as follows:

$$E(C) = \frac{\sum_{i,i>j}^{n}\log(U_{ij}-L_{ij})}{n(n-1)/2}$$

where *n* is the number of residues in the protein, L_{ij} and U_{ij} are the lower and upper bound distances between residues *i* and *j*, respectively. We set L_{ij} = 3.2 Å for all pairs, and U_{ij} = 8 Å for contacts and equal the diameter of gyration[25] DG = 5.54*$n^{0.34}$ for non-contacts.

Similar to what done in previous CASPs,[24] we evaluated the dependency between alignment depth and prediction accuracy based on the number of effective sequences in a given MSA, computed as follows:

$$\text{Neff}/L = \max(\text{Neff}_{\text{PSIBLAST}}, \text{Neff}_{\text{HHBLITS}}, \text{Neff}_{\text{BFD}}, \text{Neff}_{\text{MGNIFY}})/L$$

where $\text{Neff}_{\text{PSIBLAST}}$, $\text{Neff}_{\text{HHBLITS}}$, Neff_{BFD}, $\text{Neff}_{\text{MGNIFY}}$ are the number of effective sequences retrieved using PSIBLAST, HHBLITS, BFD and MGNIFY, respectively, and L is the length of the target.

The bin-level assessment of *distance predictions* was performed based on the average bin precision, recall, and $F1$ score, computed for each individual bin and then averaged over all distance bins, as follows:

$$\text{precision} = \frac{1}{10}\sum_{N=1}^{10} \text{TP}_N/(\text{TP}_N + \text{FP}_N)$$

$$\text{recall} = \frac{1}{10}\sum_{N=1}^{10} \text{TP}_N/(\text{TP}_N + \text{FN}_N)$$

$$F1 = \frac{1}{10}\sum_{N=1}^{10} \frac{2(\text{precision}_N)(\text{recall}_N)}{\text{precision}_N + \text{recall}_N}$$

where N in a distance bin, TP_N is the number of true positives in the bin ($p_{\max} = p_N$ and d_{ij} falls on bin N), FP_N is the number of false positives ($p_{\max} = p_N$ and d_{ij} does not fall on bin N), and FN_N is the number of false negatives ($p_{\max} \neq p_N$ and d_{ij} falls on bin N), with $p_{\max}$ being the maximum predicted probability provided over all distance bins.

For comparison, we also considered an alternative precision metric,

$$\text{precision}_{\text{overall}} = \text{TP}/(\text{TP} + \text{FP})$$

where TP is the number of true positive among all predicted residue pairs (i,j), that is, the number of pairs where $p_{\max}$ is assigned to the bin with the correct d_{ij} in the target, and FP is the number of false positives among all predicted residue pairs (i,j), i.e. the number of pairs where $p_{\max}$ is assigned to a bin which does not correspond to the correct d_{ij} in the target.

Two additional bin-level metrics were considered in the assessment of distance predictions: the mean distance difference (MDD) and the mean bin neighbor (MBN). The MDD evaluates predictions in each bin N by weighting the difference between the native and the predicted distances by the provided probability pN, as follows:

$$\text{MDD} = 1 - \left(\frac{1}{10}\sum_{k=1}^{10}\frac{1}{N_k}\sum_{a=1}^{N_k}\sum_{b=1}^{10}\frac{p_{\text{ba}}\,|\,D_k - d_b\,|}{D_{\max}}\right)$$

where N_k are the predictions falling on the kth bin, p_{ba} is the probability assigned to the bth bin in the ath prediction, $|D_k - d_b|$ is the difference between the observed distance and the mean distance of the bth bin, and $D_{\max}$ is 21 Å.

The MBN evaluates predictions by summing the probability assigned to the bin where the observed distance falls with those assigned to the two neighboring bins, adjusted by a factor of 0.5, as follows:

$$\text{MBN} = \frac{1}{10}\sum_{k=1}^{10}\frac{1}{N_k}\sum_{b=1}^{N_k}\left(p_b(D_k) + \frac{p_b(d_{k-1}) + p_b(d_{k+1})}{2}\right)$$

where N_k are the distances falling on the kth bin, $p_b(D_k)$ is the probability assigned to the bin where the observed distance falls (D_k), and $p_b(d_{k-1})$ and $p_b(d_{k+1})$ are the probability assigned to the two neighboring bins of D_k.

Additionally, distance predictions were assessed using graph-based metrics as described below. This was intended as a way of integrating into the evaluation the global properties of the predicted distance maps, and, at the same time, assess the contribution of each individual residue to the overall prediction accuracy. Each distance map was represented as an undirected, weighted graph with amino acid pairs i and j as nodes and weights as edges. Weights are defined as $w_{ij} = 1/d^2$, where d is the mean distance in the $p_{\max}$ bin. If $p_{\max} = p_{10}$, the distance d was assigned a fixed value of 21 Å. If more than one bin was predicted with equal $p_{\max}$, a single bin was randomly selected and its corresponding distance range considered in the assessment. Similarly, given the distance map corresponding to the native protein structure, a graph was built by defining the edge weight between nodes i and j as $\text{w}_{\text{ij}} = 1/d^2$, where d is the observed distance between the C_β atoms of the amino acid pair (C_α in case of Glycine). When the distances in the native structure were greater than 20 Å, d was set as the average value of all distances > 20 Å for that target. Subsequently, for each graph we computed the following parameters:

$$s_i = \sum_{j=1}^{n} a_{ij} w_{ij}$$

where the s_i is the strength local parameter[26] of the ith residue, a_{ij} are the elements of the adjacency matrix, and s_i is computed as the sum of the weights of the adjacent edges j.

The clustering coefficient local parameter[26,27] of the ith residue is defined as:

$$c_i^w = \frac{1}{s_i(k_i - 1)}\sum_{j,h}\frac{(w_{ij} + w_{ih})}{2}a_{ij}a_{ih}a_{jh}$$

where s_i and k_i are the strength and degree of residue i, respectively, and a_{ij} (as well as a_{ih} and a_{jh}) are the elements of the adjacency matrix and w_{ij} (and w_{ih}) are the weights.

The normalization factor $s_i(k_i - 1)$ ensures that $0 \le c_i^w \le 1$.

The average shortest path global parameter[28] of the ith residue is defined as:

$$\underline{sp_i} = \frac{1}{n}\sum_{j=1}^{n} sp_j$$

and each shortest path connecting two nodes of the graph is defined as the path that minimizes the sum of a given real-valued weight function:

$$sp_{i\to j} = P(v_1,\ldots,v_i,\ldots,v_n) \mid P = \min\left(\sum_{i=1}^{n-1} f(w_{i,i+1})\right)$$

where i and j are the residues, $sp_{i\to j}$ is the shortest path between nodes i and j as computed by the Dijkstra algorithm,[29] n is the total number of nodes and w is defined as the inverse value of the weight.

The diversity global parameter of the ith residue $D(i)$, as defined in,[30] is the normalized entropy of the weights of the normalized weights of all edges departing from a given node, calculated as follows:

$$D(i) = \frac{H(i)}{log(k_i)}$$

where

$$H(i) = -\sum_{j=1}^{k_i} p_{ij} log(p_{ij})$$

is the Shannon entropy of the ith residue, and

$$p_{ij} = \frac{w_{ij}}{\sum_{l=1}^{k_i} w_{il}},$$

where w_{ij} is the weight of the edge between residues i and j, k_i is the degree of node i, and l runs over all neighbors of node i.

The comparison between predicted and native graphs was based on the Pearson correlation for all metrics. To rank the participating groups according to their performance both in distance and contact assessment, all metrics were transformed into z-scores. The per-target z-score of a group was set to zero if they did not submit a prediction on a given target. Finally, the cumulative rank of a group was assigned based on the sum of its per-target z-scores greater than zero.

3 | RESULTS

3.1 | Performance comparison to previous CASPs

Figure 1 shows the results of the participating methods during the latest 4 rounds of contact prediction in CASP. In order to facilitate the comparison to previous CASP assessments, results of the CASP14 contact prediction assessment are reported in terms of average precision for FM domains and the $L/5$ lists of long-range contacts, unless specified otherwise. On average, the top 5 predictors in CASP14 (tFold-CaT_human, tFold-IDT_human, TripletRes, PreferredFold and DeepPotential) achieve 64% precision (22 domains), a similar performance as observed among the top 5 groups (RaptorX-Contact, TripletRes, ResTriplet, GREMLIN_baseline and TripletRes_AT) in CASP13 (65% precision on 31 domains). While these results may indicate a setback in the advancement of contact prediction methods, we should emphasize that progress in this round might be offset by the increased difficulty of the CASP14 targets. As it can be seen in Figure 2A, CASP14 FM targets have the lowest coverage and sequence identity to available structural templates compared to all previous CASPs. This is also reflected by the number of effective sequences, which shows that 30% of the FM targets in CASP14 have very small numbers of homologous sequences (Neff/L < 0.2), while this was only the case for about 10% of the targets in the CASP13 FM dataset (Figure 2B). Interestingly, for targets with similar Neff/L, it is possible to see that the maximum achieved precision per target is higher in CASP14 vs CASP13 (Figure 5). On average, maximum achieved precision per target (FM) reaches 55% in CASP14, compared to 50% in CASP13. Notably, CASP14 best performance per target compares favorably both in the low (Neff/L < 0.2) and high range of Neff/L values (Neff/L ≥ 0.2), with 38% and 61% average best precision respectively, to be compared with 29% and 52% achieved in the previous round. Notably, the extended FM + FM/TBM target set in CASP14 is only comparable in terms of sequence identity to CASP13 FM targets, with the difference that the CASP14 target set lies in a

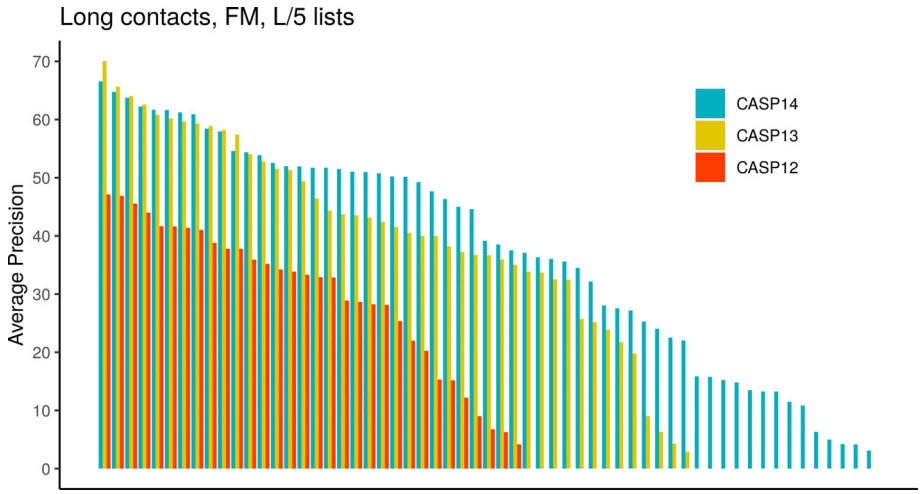


FIGURE 1 Improvement of contact prediction over CASP11-CASP14 meetings. Participating groups (X-axis) are ranked according to average precision (Y-axis) for the top $L/5$ contacts

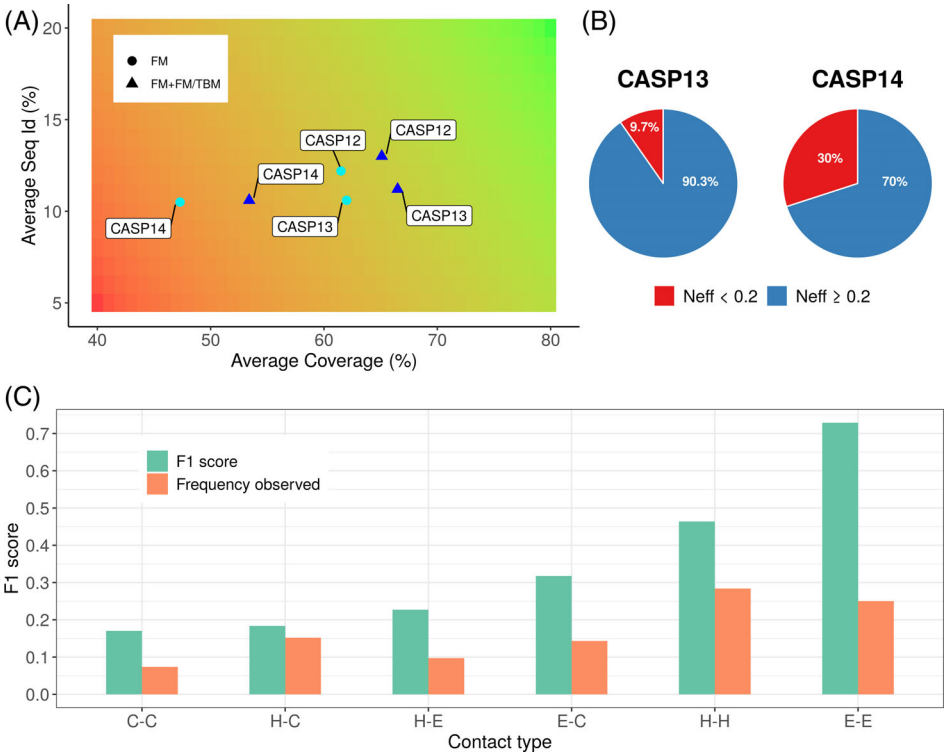


FIGURE 2 Analysis of target difficulty in CASP14. (A) Scatter plot representing the average sequence identity vs. average coverage of the best structural templates from CASP12 to CASP14. Cyan circles correspond to FM targets while blue triangles represent FM + FM/TBM targets. Red to green gradient of the plot box reflects predictive difficulty from harder to easier, with the lower bottom corner hosting the most difficult target sets. (B) Proportion of FM targets with low alignment depth (Neff < 0.2) in CASP13 vs CASP14. Data are only shown for targets containing long-range contacts. (C) Contact predictions accuracy as a function of connected secondary structure elements (*x*-axis). Results are shown in terms of average *F*1-score (green bars) for long-range *L*/5 contacts (all participating groups). Orange bars indicate the overall frequency of long-range contacts as a function of connected secondary structure elements in target structures

much lower range of sequence coverage (Figure 2A). In this regard, it is worth noting that CASP14 methods compare favorably on this target set with an average precision of 74% for the top five groups vs 64% seen in the previous round (Figure S1).

We analyzed the accuracy of contact prediction with respect to the type of secondary structure elements, which were extracted from the experimental structures using the DSSP program.[31,32] About 37% of the long-range contacts present in the FM target dataset involve at least one residue from a coil element, 53% are alpha-helices mediated contacts, and about 47% are beta-strands mediated contacts (Figure 2C). In terms of accuracy, β-strand mediated contacts are generally predicted with markedly higher accuracy, especially β-β contacts (*F*-score = 0.7), compared to both alpha helices and coil mediated contacts (Figure 2C). The latter remains most challenging for predictors, with average *F*-score below 0.2 for all contact types. While the results of this analysis are overall in line to what observed in CASP13,[20] it is worth highlighting that in terms of secondary structure content the two sets of FM targets show a different composition, with CASP14 FM targets showing an overall lower content of β-strands compared to CASP13 FM targets (Figure S2).

3.2 | Assessment of contact predictions

The results of the analysis of group performance for long range contacts in *L*/5 contact lists is shown in Figure 3. For each group, results are shown in terms of cumulative *z*-scores based on the *F*1 + 0.5*ES (ext) metric computed over FM targets. Overall, the top 10 groups achieve comparable performances, with an average upper limit of precision of ∼66% (Figure 1). A similar ranking of the top performing groups is observed across different contact ranges. Specifically, groups G368 (tFold-CaT_human) is the top ranked group in all rankings, followed by G488 (tFold-IDT_human), G024 (DeepPotential) and G009 (tFold_human) among the top 5 groups in both medium+long and short-range contacts (Figure S3). Likewise, the ranking solely based on the precision is very similar to the adopted combined ranking, with the top three groups being the same in the same order (368, 10, 488) and the top 10 groups being the same in slightly shuffled order (Figure S4).

A head-to-head comparison performed based on common target domain sets (Figure 4) did not reveal statistically significant differences in the per-target performance between the top 10 groups with the exceptions of methods G368, G488 and G009 (tFold-CaT_human,

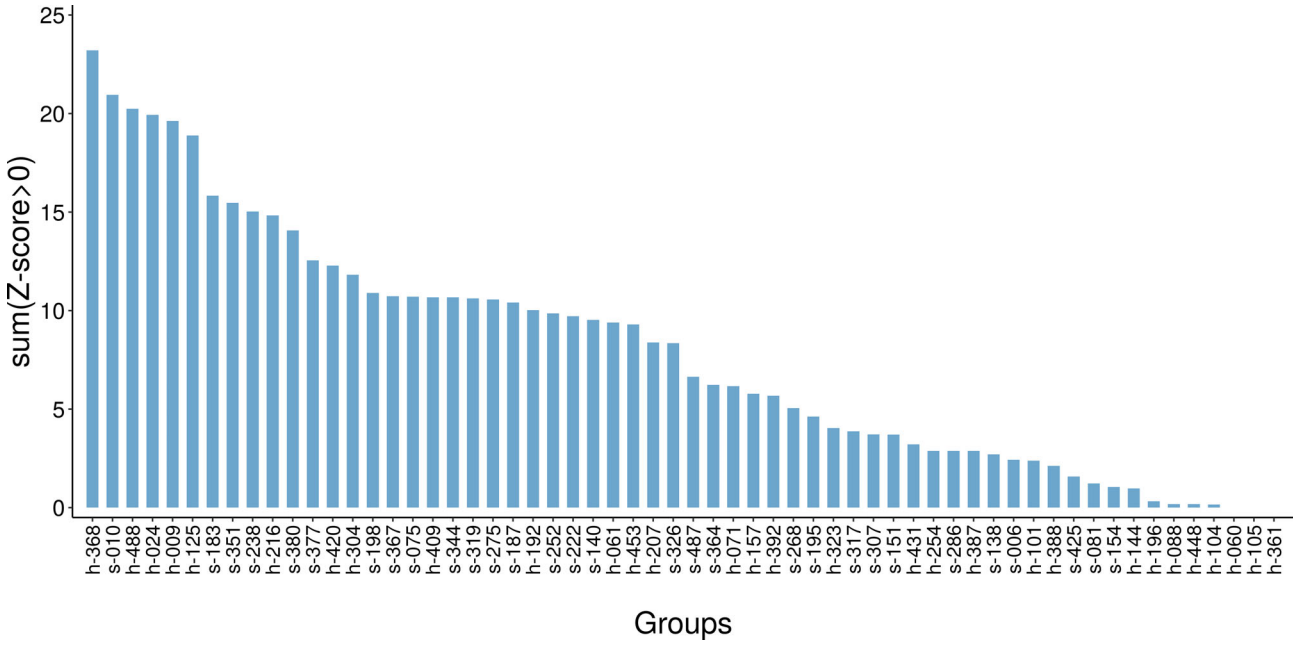
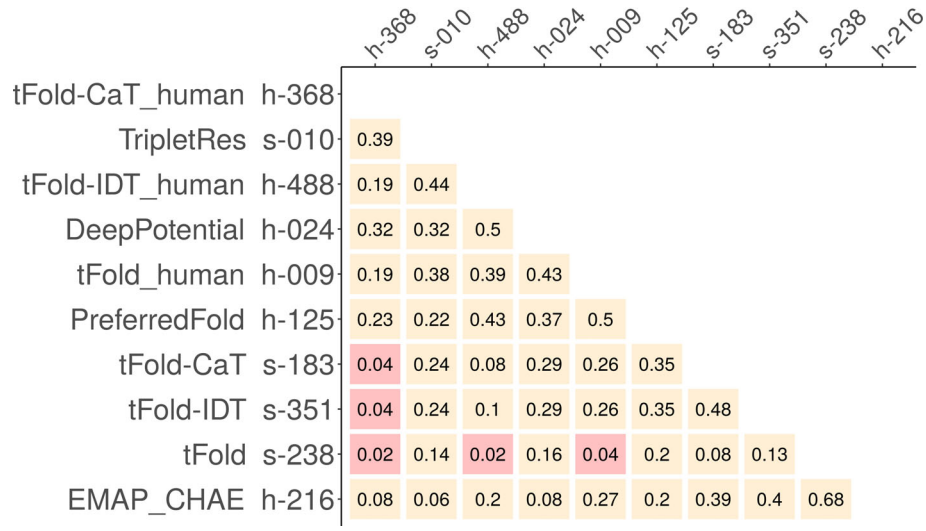


FIGURE 3 Cumulative *z*-score ranking of participating groups on FM targets. Performance is shown for the top *L*/5 long-range contacts. Group names are labeled as *h* and *s* to denote human-expert and server methods, respectively

FIGURE 4 Results of the paired Student's *t* test computed on the top-10 performing groups according to cumulative *z*-score ranking. Red cells correspond to *p*-value < .05



tFold-IDT_human, tFold_human), which significantly outperform methods G183, G238 and G351 (tFold-CaT, tFold and tFold-IDT). Interestingly, these two sets of methods are developed by the same research group (Tencent AI). Two main differences among these methods are the choice of input features used to feed the deep-learning neural network (e.g., multi-MSA ensembles and 2D attention modules in the case of tFold-Cat methods vs single MSA and template based features in the case of tFold-IDT methods), and the regime of generating predictions (as automatic servers with models due in 3 days, or human-expert groups with extra 18 days for the refinement step). Apparently, the refinement step, which embeds additional information from structural decoys, had a major impact on the accuracy of the Tencent models.

Finally, it is worth mentioning that the TripletRes (G010) server achieved comparable performance to the best human-expert methods, with average precision on FM targets of 64% and 71% when considering top *L*/5 long and medium+long range contacts, respectively, and reaching 80% when considering the FM + FM/TBM targets on the same contact ranges (Figure S1).

3.3 | Prediction performance as a function of alignment depth

The use of coevolution data has been the main driver of the observed improvements in contact prediction during previous CASPs.[24] As

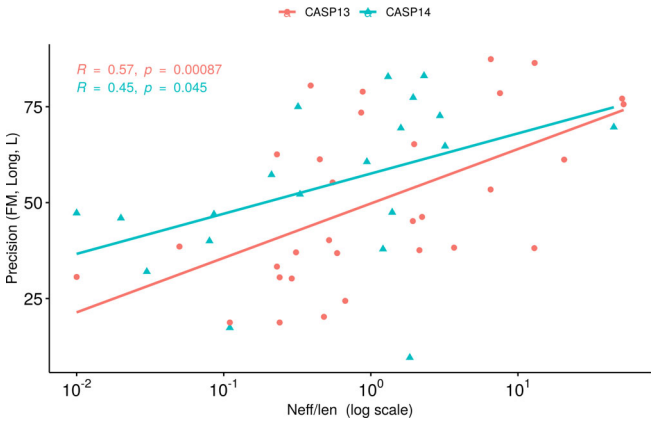


FIGURE 5 Contact precision of the best prediction method as a function of alignment depth. Data are shown for Top *L* long-range contacts and FM targets in CASP13 and CASP14. R refers to Pearson coefficient and *p* to *p*-value

coevolution-based features strongly depend on the availability of homologous sequences,[33,34] the effective depth of MSAs is generally considered to be a determining factor for the accuracy of contact predictions. Such a relationship was apparent in CASP12[24] and CASP13,[35] and it is still noticeable in the present round ($R = 0.45$)(Figure 5).

Remarkably, about 30% of the CASP14 FM targets had very little (Neff/L < 0.2) or essentially non existing sequence information (Neff/$L = 0.01$) (see Methods).[20,24] Nonetheless, remarkable precision was achieved for some targets, such as T1043-D1 and T1074-D1, with precision of 0.97 and 0.96, respectively. In general, predictions where highly accurate for targets with relatively high number of homologous sequences (Neff/L > 0.3, average precision = 88%), with two exceptions, such as T1029-D1 (Neff/$L = 1.84$) and T1064-D1 (Neff/$L = 0.11$), which turned out to be very challenging targets for all predictors, with maximum achieved precision on Top *L* contacts of 10% (20% based on Top L5) and 17% (13% based on Top L5), respectively.

In summary, the top performing group benefited from large metagenomics libraries and the depth of the MSA. Performance is also affected by the length of the target, although to a lower extent which is only noticeable for FM targets (Figure S5). TripletRes and DeepPotential, in particular, rely on using BFD/MGnify for sequence search. On the other hand, while the tFold family of methods did not rely on the same large metagenomics databases, they reached top performance by leveraging large ensembles of MSAs alignments. The different performance achieved by two very similar algorithms, that is, G453 (DMP2) and G304 (Jones-UCL), highlights the importance of the MSA generation step, where the manual curation of the alignments constitute the main difference between the two methods (Figure 3). Finally, as anticipated in CASP13, the application of deep neural networks seems to be a key factor as most of the top performing groups based their methods on this framework, leading to high prediction accuracy even in the near absence of evolutionary information.

3.4 | Assessment of distance predictions

Distance predictions were assessed based on two different sets of evaluation metrics, as described in the Methods section. Figure 6 shows a comparison of performance in terms of *z*-score for all metrics and considering the full FM + FM/TBM target set. Clustering was performed based on complete linkage of Euclidean distances and results visualized as heatmap colored from yellow to blue, indicating low to high performance respectively. As it can be seen from the left-side dendrogram, the analysis identified four main clusters of participating groups. A first cluster is composed of 5 groups (G010, G024, G368, G488, and G009) achieving top performance according to all metrics. Within this cluster, methods are further grouped by research groups of origin, that is, methods G024 (DeepPotential) and G010 (TripletRes) from the Zhang lab, and methods G368 (tFold-CaT_human), G488 (tFold-IDT_human), and G009 (tFold_human) from the Tencent AI lab. Although marginal, differences in performance between the two subclusters are captured by the graph-based metrics, where predictions from the Tencent AI seem to better capture interaction hubs (clustering coefficient) and patterns of long-term distances and inter-residue distance distributions (shortest path and diversity). Groups G351, G183, G192 and G304 constitute a second main cluster with lower performance. Within this cluster, a further segregation can be seen between human-expert methods (G304 and G192) vs server methods (G351 and G183), as revealed by bin-level metrics as well as local topological graph-based metrics, i.e. strength and clustering coefficient. As it can be seen from the top dendrogram, the clustering analysis segregates the bin-level metrics (*F*1-score, MDD, and MBN) and graph-level metrics (diversity, strength, clustering coefficient, and shortest path) into separate branches, reflecting their different nature. Despite their differences, the resulting rankings of top performing groups across all metrics are largely consistent (Figure S6).

In terms of precision, the top 5 predictors achieve 32% average bin precision on long-range distances if considering FM domains (-Figure S7) and 38% on FM + FM/TBM domains. These results refer to the assessment on full distance maps and thus are expectedly lower than the contact precision, which is calculated on the subsets of the most reliably predicted contacts. Additionally, analysis of the dependency of the average bin precision on the depth of alignment showed no correlation between the two measures (Figure S8).

In order to calibrate precision of long-range distances versus 3D model accuracy, we correlated the *predicted* precision and GDT_TS scores of 3D models from groups who participated in both prediction categories. The correlation appeared to be low (Pearson CC = 0.47), and thus it was hard to reliably establish the dependency of expected 3D model accuracy on the precision of distance prediction. To approach the problem from a different perspective, we analyzed the dependency of the bin precision of long-range distances *extracted from 3D models* submitted to the TS category versus GDT_TS scores of the models. In this case, the two values correlated well (Pearson CC = 0.9). Figure S9 shows that the average bin precision of 32%, which was achieved by the best distance prediction groups,

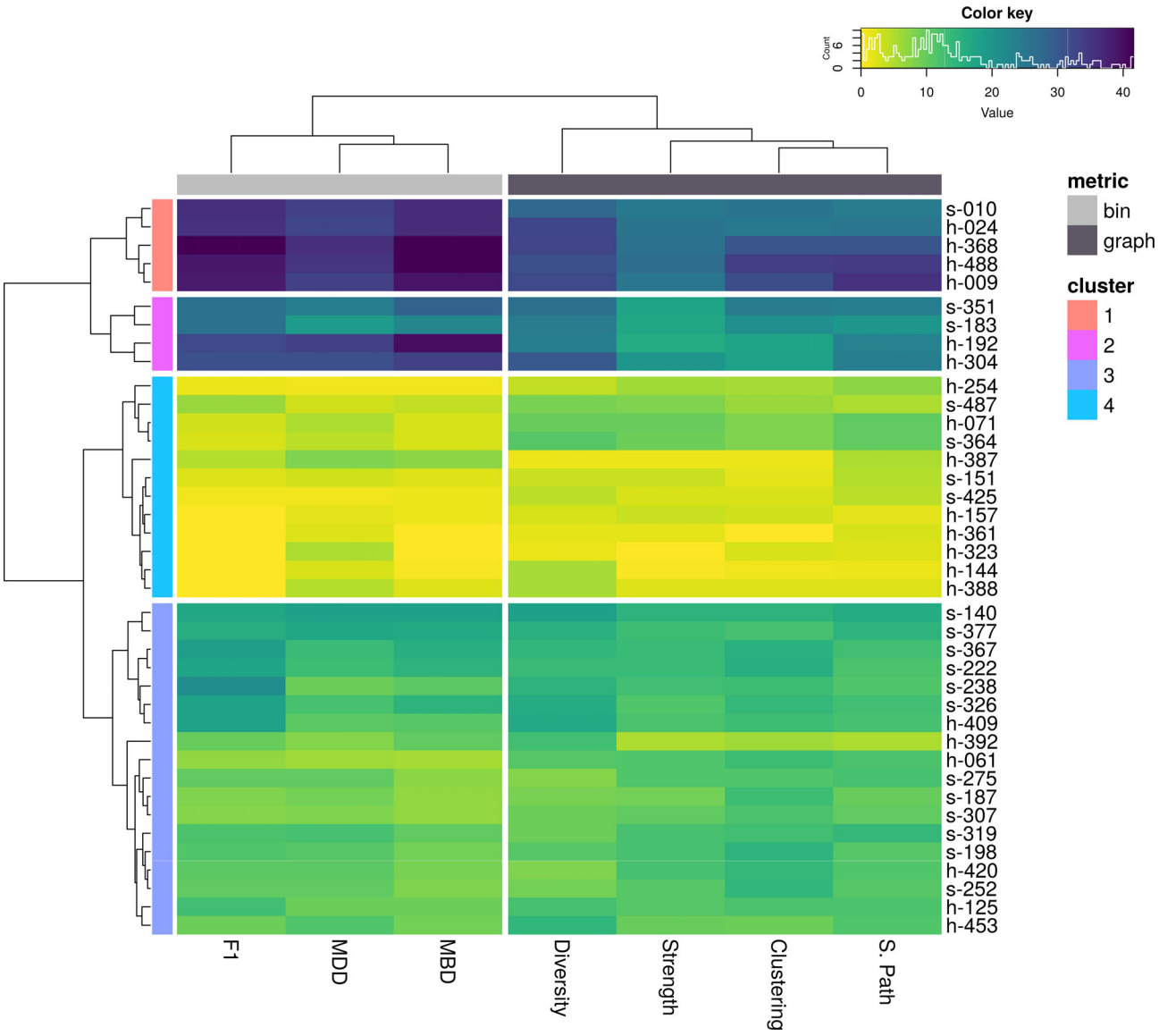


FIGURE 6 Heatmap clusters of group performances and assessment metrics. Columns include four bin-level assessment scores (average bin precision, *F*1, MDD, and MBN) and four graph-based scores (diversity, strength, clustering coefficient, and shortest path). Rows represent participant groups. Rows and columns were clustered using Euclidean distance with complete linkage. The heatmap is colored according to the per-group sum of *z*-score values >0 for each metric, and ranges from yellow (low) to blue (high)

corresponds to the expected GDT_TS of 50. This indirectly suggests that the accuracy of distance predictors in CASP14 is still insufficient to achieve 3D results comparable with those from the AlphaFold2 group,[37,38] who attained GDT_TS scores in excess of 80 for 90% of targets (GDT_TS of 80–85 corresponds to the expected long-range distance precision of 55%). Indeed, although the AlphaFold2 (G427) did not participate in the distance prediction category, it is worth noticing that by assessing the distances extracted from AlphaFold2 models, top average precision on the set of FM targets reaches 62%, almost doubling the performance of the second best predictor on this dataset (Figure S10).

We have also considered an alternative precision metric (overall precision, see Methods), which takes into account the overall number of TPs and FPs (one per predicted pair) in the whole prediction. As expected, this metric leads to a higher average precision over all groups, with top 5 predictors achieving 66% on both FM (Figure S11A) and FM + FM/TBM domains. However, the score is heavily affected by the excess of long-range distances in the last distance bin (>20 Å), leading to a poor discrimination capacity of this metric in terms of group performance as well as accuracy of corresponding 3D models (Pearson CC with GDT_TS = 0.4, Figure S12A). However, when only the first nine bins are considered, the correlation with GDT_TS increases to 0.9 while the overall precision drops to 37% for FM and 39% for FM + FM/TBM (Figures S11B and S12B), in line with the results of the per-bin evaluation (Pearson CC = 0.9, Figure S13).

With the aim of unifying the results into a single ranking, a meta-score was designed by combining all the individual scores that focus on different features of the predictions. To this aim, we removed possible biases towards any evaluated aspect by discarding metrics that shared high correlation values (R > 0.80, Figure S14). The final score was

defined as the linear combination of 5 nonredundant metrics (*z*-MDD + *z*-*F*1 + *z*-clustering + *z*-diversity + *z*-shortest_path) and the results of the corresponding group ranking are shown in Figure 7. Additionally, a head-to-head paired *t* test on common target domains was performed across the top-10 groups according to the metascore (Figure 8). Results indicate that there is no statistical difference between the top-5 performing groups (G368, G488, G010, G009 and G024). Significant differences are observed between the top-3 groups (G488, G368 and G009), which significantly outperform all other groups from rank 6, between G010 (TripletRes) and G140 (YangServer), as well as between G351 (tFold-IDT) and G183 (t-Fold-CaT).

Notably, although the assessment of contact and distance predictions differed in terms of both evaluation metrics and assessed lists, that is, *L*/5 vs full distance maps, the same five groups G368, G488, G010, G009 and G024 were identified as top ranking in both cases.

3.5 | Graph-based metrics in detail

In this section, we provide a description of the results of the assessment on two targets with the aim of illustrating how different assessment metrics behave, with a special emphasis on the graph-

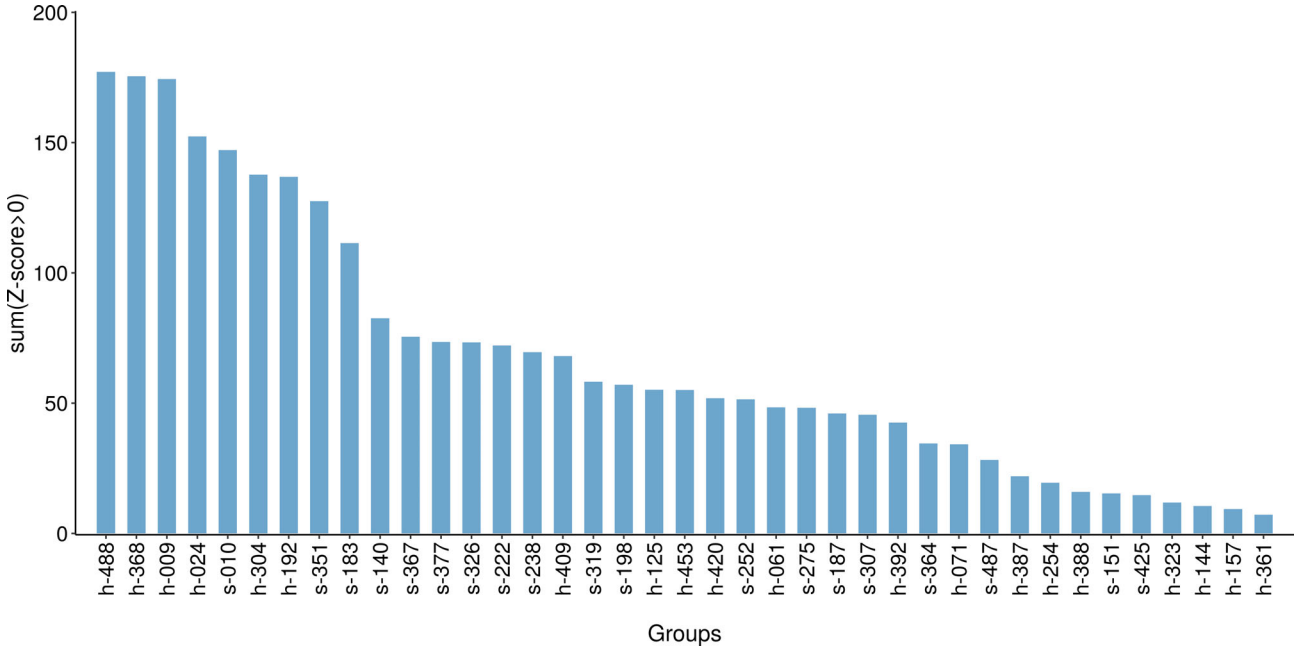


FIGURE 7 Cumulative *z*-score ranking of participating groups in the distance prediction category. Performance is shown according to the linear combination of five nonredundant metrics (*z*-MDD + *z*-*F*1 + *z*-clustering + *z*-diversity + *z*-shortest_path). Group names are labeled as *h* and *s* to denote human-expert and server methods, respectively

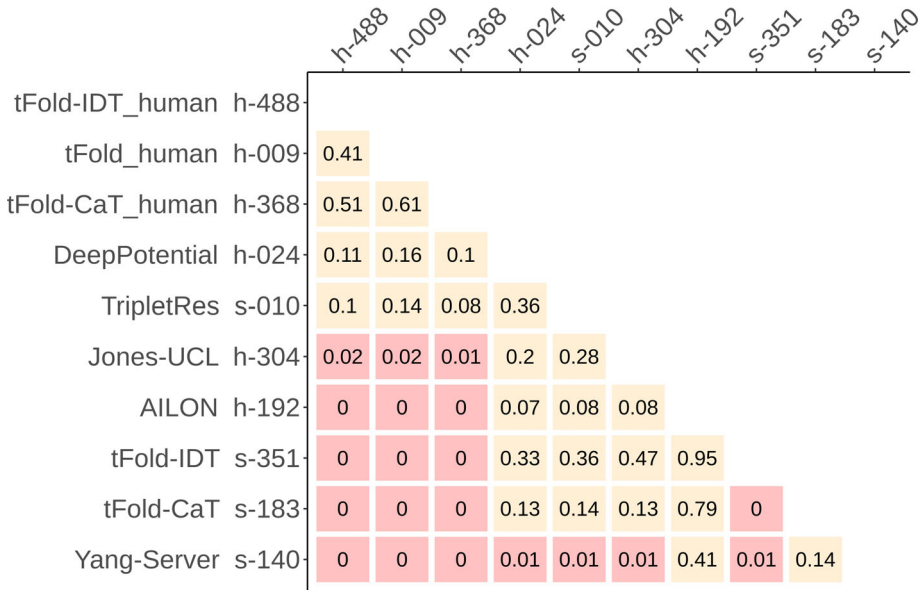


FIGURE 8 Results of the paired Student's *t* test computed on the top-10 performing groups according to the metascore-based ranking. Red cells correspond to *p*-value < .05

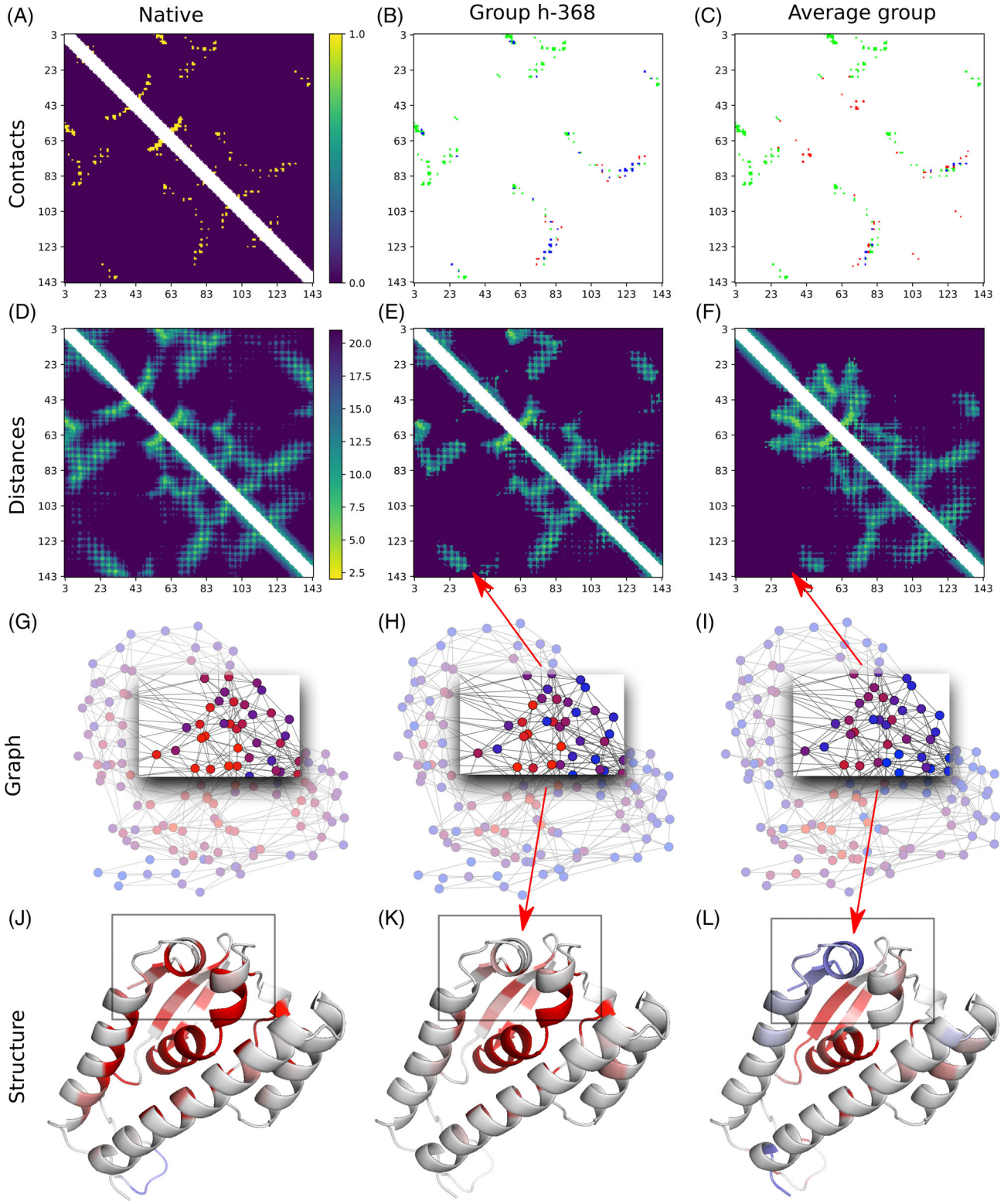


FIGURE 9 Target T1093-D1, log(Neff/len) = 0.11. (A–C) Contact maps. (D–F) Distance maps. (G–I) Graphs-based representation of predicted and native distance maps. Graph nodes represent amino acid residues and are colored based on per-residue strength values. (J–L) 3D structure of the T1093-D1 target colored according to per-residue strength values. Color ranges from red (high strength) to blue (low strength). Highlighted regions correspond to amino-acid residues showing the highest strength values in the native graph

based metrics. In the first example, we show a comparison of the results obtained in the assessment of contact and distance predictions by two groups: a first group, achieving top performance according to *F*1 score, and a second group, reflecting the average performance achieved from all participants. In the second example, target T1080-D1 is chosen to highlight how different graph-based

metrics capture local and global topological features of the predicted distance maps, and how they reflect different structural features of the target.

Figure 9 shows the results on target T1093-D1. The top performing group for this target, G368 (tFold-CaT_human), achieved high accuracy results both in contact and distance predictions (*F*-score = 0.368 and 0.329, respectively), despite the target was generally very challenging for most groups (average *F*-score contacts = 0.123; average *F*-score distances = 0.154). Predicted contact maps are shown in panels B and C and correspond to long-range *L*/5 contacts, where green dots refer to long-range contacts of the target, blue dots correspond to correct predictions (TP) and red dots to incorrect predictions (FP). As it can be seen in panel B, G368 identified contacts that are distributed across four main contact hubs in the targets. Both groups predicted contacts in a hub of alpha-helices (range 75–85/120–130), where G368 predicted mostly TP contacts. The average predictor showed lower accuracy in this region and also over-predicted two additional hubs, 43–48/70–75 and 103–106/127–130, that are absent from the target. These results are also reflected in the predicted distance maps showing that group G368 reproduced correctly different patterns of distances observed at the N- and C-termini of the target, while the average predictor missed them and, instead, over-predicted at the level of the β-sheet (Figure 9E,F). This translates into differences in local and global topological features of the corresponding prediction graphs, as shown for the strength local parameter (Figure 9H,I). The per-node strength correlation between predicted graph and native graph is indeed high for group G368 ($R = 0.88$), but low for the average performing group ($R = 0.4$). Mapping the predicted node strength values (i.e., the sum of the weights of the edges departing to/from a given node) onto the protein structure (Figure 9J) shows that G368 predicts a similar pattern to that seen in the target at the level of both α-helical elements and β-sheets

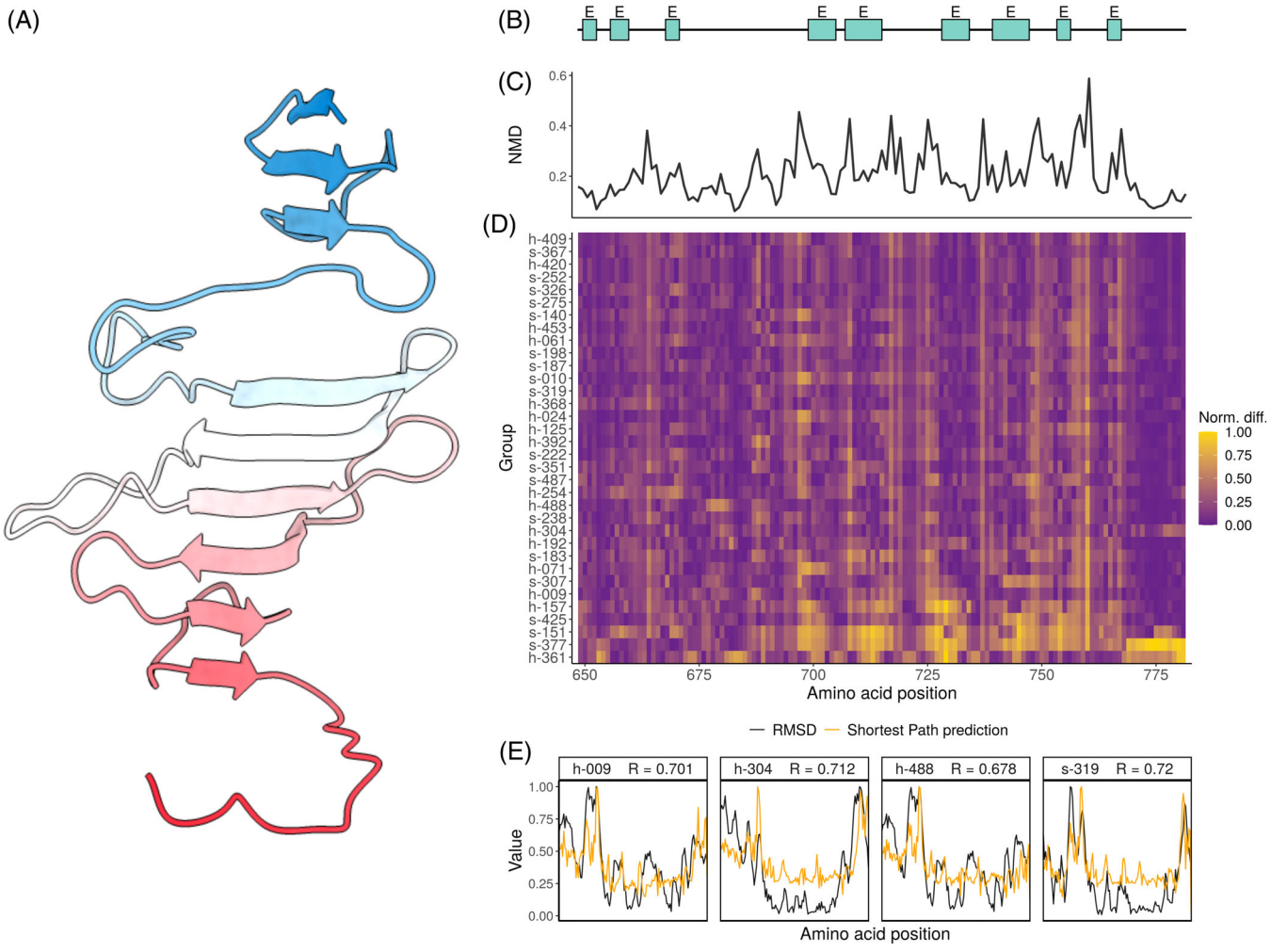


FIGURE 10 (A) 3D structure of target T1080-D1, N-terminal and C-terminal in blue-red gradient color. (B) Diagram of secondary structure content. β-strand elements are indicated by green rectangles while turns and loops are denoted by the black line. (C) Normalized mean difference (NMD) between observed and predicted per-residue strength computed over all predictions. (D) Heatmap showing the normalized absolute difference in per-residue strength between predicted and native graphs for each participant group (*y*-axis). (E) Absolute difference between predicted and native per-residue shortest path (Δs_path, yellow line) and RMSD (black line) for predictors G009, G304, G488 and G319. Values of Δs_path and RMSD are shown as normalized values between 0 and 1

(Figure 9K), while the average predictor exhibits major differences in both secondary structure elements (Figure 9L).

Graph-based metrics were adopted in order to capture both the local features of the predicted distance maps (strength and clustering coefficient) as well as to characterize global patterns of the molecular interaction network (average shortest path, that is, the average length of all shortest paths departing from/to a given node, and diversity, that is, the scaled Shannon entropy of the weights of the edges departing to/from a given node). Figure 10 illustrates how different metrics reflect different features of the predictions in the structural context of the target. The shown target (T1080-D1) is 133 amino acids long, composed of 9 β-sheets connected by both turns and long loops (Figure 10A,B). These loops correspond to regions of the predicted graphs where the per-node strength values disagree the most with those of the native graph (Figure 10C,D). Figure 10E, in turn, shows the difference between the average shortest path global parameter (orange line) computed for each node of the predicted graph from four predictors (G024, G252, G368 and G420) versus the same parameter computed on the native graph (Δs_path). In this case, major deviations from the native graph are observed at the N- and C-terminal of the protein (Figure 10E).

As contact and structure predictions are submitted separately in CASP, it is difficult to analyze how the accuracy of contact predictions influences the 3D model accuracy. However, for those cases where participants submitted in both categories, it is worth observing where relationships exist or fail to.[20] In the case at hand, a high correlation exists between the Δs_path and the RMSD of the 3D model submitted by the same prediction group ($R = 0.7$ on average, Figure 10E). For both metrics, lower deviations from the native are observed at the level of the core antiparallel β-sheet motif of the target, while larger errors in both the predicted distance maps and three models are observed in the protein terminal segments. In conclusion, the different graph-based metrics presented here provide a sensitive way to assess distance predictions, enabling a deeper understanding of both the local and global characteristics of the predicted distance maps.

4 | CONCLUSIONS

Over the years, contact prediction has evolved from being the niche of specialized groups[3] to one of the clearest examples of theoretical and methodological advancements in the history of CASP.[20,24] With regards to performance of the methods in CASP14, the results of the assessment in contact prediction did not reveal a discernible progress when compared to the previous CASP round. Average precision in CASP14 reached 64%, slightly below the limit observed in CASP13. Arguably, this could be due to CASP14 FM targets being more challenging than ever before. At the same time, a larger number of participants in CASP14 reached an average precision of 50%, which is suggestive of continuous development and advances in the field. Contact prediction remains as an attractive challenge for the community with more than 60 groups participating in the category (30% increase compared to CASP13). Participation was also remarkable in the new category of inter-residue distance prediction, which attracted 39 groups. As opposed to contact prediction assessment, which was based on the standard protocol adopted in previous CASPs, distance prediction assessment required the development of a novel, ad-hoc procedure. In particular, the assessment relied on different sets of metrics, which evaluate distance pairs individually as well as in the context of the whole distance network, using a graph-based analysis framework. For long-range distances and FM targets, the top ranking methods reached 32% precision on the average and 64% for the most accurately predicted contacts. In summary, despite the different formats, metrics and procedures adopted in contact and distance predictions, the results of the assessment indicate that predictions submitted by Tencent AI lab (tFold framework) and Zhang lab (TripletRes and DeepPotential) were the most accurate in both categories.

The main differences between top performing methods stem from the MSA generation/selection step rather than the distance prediction itself, which is mostly done by deep neural networks (DNN) with a further quality refinement step. TripletRes and DeepPotential, in particular, leverage large metagenomics databases, such as MGnify and BFD, to build a few candidate MSAs which are used to produce different distance predictions using DNNs. The tFold family of methods, on the other hand, do not rely on large metagenomics databases but have a distinctive feature: starting from an ensemble of many different MSAs, distance predictions are made using DNNs equipped with attention modules, clustered into distinctive patterns, and finally the best prediction is selected based on quality assessment and second round of clustering. With these large ensembles of MSAs, tFold is able to reduce the amount of noise in the predictions and achieve impressive results. Jones-UCL (G304), another top-performing method in distances, shows significantly improved performance on a very similar prediction pipeline (DMP2, G453) by including manual steps, such as manual domain parsing, assembly of multi-domain models and alternative alignments. This significant difference in performance achieved by different methods highlights the importance of the MSA generation step for prediction. In conclusion, after the dramatic progress seen in de novo protein structure prediction in CASP14, we are looking forward to seeing if in future CASPs, contact prediction will still remain a necessary task of protein structure prediction algorithms, or whether inter-residue distance prediction will establish itself as the core step, as the state-of-the-art in this CASP is hinting at.

ACKNOWLEDGMENT

This research was supported by the US National Institute of General Medical Sciences (NIGMS/NIH) grant GM100482 (AK).

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/prot.26248.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the CASP website at https://predictioncenter.org/casp14.

ORCID

Victoria Ruiz-Serra https://orcid.org/0000-0003-3991-0514
Camila Pontes https://orcid.org/0000-0002-8726-3641
Edoardo Milanetti https://orcid.org/0000-0002-3046-5170
Andriy Kryshtafovych https://orcid.org/0000-0001-5066-7178
Rosalba Lepore https://orcid.org/0000-0002-9481-2557
Alfonso Valencia https://orcid.org/0000-0002-8937-6789

REFERENCES

1. Göbel U, Sander C, Schneider R, et al. Correlated mutations and residue contacts in proteins. *Proteins*. 1994;18:309-317.
2. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. *Nat Rev Genet*. 2013;14:249-261.
3. Lesk AM. CASP2: report on ab initio predictions. *Proteins*. 1997;29 (S1):151-166.
4. Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des*. 1997;2:S25-S32. 10.1016/s1359-0278(97)00060-6
5. Havel TF, Crippen GM, Kuntz ID. Effects of distance constraints on macromolecular conformation. II. Simulation of experimental results and theoretical predictions. *Biopolymers*. 1979;18:73-81. 10.1002/bip.1979.360180108
6. Brünger AT, Clore GM, Gronenborn AM, et al. Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: application to crambin. *Proc Natl Acad Sci U S A*. 1986;83:3801-3805.
7. Clore GM, Nilges M, Brünger AT, et al. A comparison of the restrained molecular dynamics and distance geometry methods for determining three-dimensional structures of proteins on the basis of interproton distances. *FEBS Lett*. 1987;213:269-277.
8. Orengo CA, Bray JE, Hubbard T, et al. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins*. 1999;37(S3):149-170.
9. Lesk AM, Lo Conte L, Hubbard TJP. Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts. *Proteins: Struct Funct Genet*. 2001;45:98-118. 10.1002/prot.10056
10. Aloy P, Stark A, Hadley C, et al. Predictions without templates: new folds, secondary structure, and contacts in CASP5. *Proteins*. 2003;53-(suppl 6):436-456.
11. Graña O, Baker D, MacCallum RM, et al. CASP6 assessment of contact prediction. *Proteins*. 2005;61(suppl 7):214-224.
12. Izarzugaza JMG, Graña O, Tress ML, et al. Assessment of intramolecular contact predictions for CASP7. *Proteins*. 2007;69(suppl 8):152-158.
13. Ezkurdia I, Graña O, Izarzugaza JMG, et al. Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. *Proteins*. 2009;77(suppl 9):196-209.
14. Monastyrskyy B, Fidelis K, Tramontano A, et al. Evaluation of residue-residue contact predictions in CASP9. *Proteins*. 2011;79(suppl 10):119-125.
15. Monastyrskyy B, D'Andrea D, Fidelis K, et al. Evaluation of residue-residue contact prediction in CASP10. *Proteins*. 2014;82(suppl 2): 138-153.
16. Monastyrskyy B, D'Andrea D, Fidelis K, et al. New encouraging developments in contact prediction: assessment of the CASP11 results. *Proteins*. 2016;84(suppl 1):131-144.
17. Kosciolek T, Jones DT. Accurate contact predictions using covariation techniques and machine learning. *Proteins*. 2016;84(suppl 1):145-151.
18. Wang S, Sun S, Xu J. Analysis of deep learning methods for blind protein contact prediction in CASP12. *Proteins*. 2018;86(suppl 1):67-77.
19. Buchan DWA, Jones DT. Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins*. 2018;86(suppl 1): 78-83.
20. Shrestha R, Fajardo E, Gil N, et al. Assessing the accuracy of contact predictions in CASP13. *Proteins*. 2019;87:1058-1068.
21. Xu J, Wang S. Analysis of distance-based protein structure prediction by deep learning in CASP13. *Proteins*. 2019;87:1069-1081.
22. Li Y, Zhang C, Bell EW, et al. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. *Proteins*. 2019;87:1082-1091.
23. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature*. 2020;577:706-710.
24. Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, et al. Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. *Proteins*. 2018;86(suppl 1):51-66.
25. Kryshtafovych A, Fidelis K, Moult J. CASP10 results compared to those of previous CASP experiments. *Proteins*. 2014;82(suppl 2): 164-174.
26. Barrat A, Barthelemy M, Pastor-Satorras R, et al. The architecture of complex weighted networks. *Proc Natl Acad Sci U S A*. 2004;101: 3747-3752. 10.1073/pnas.0400087101
27. Wasserman S, Faust K. *Social Network Analysis: Methods and Applications*. Cambridge University Press; 1994.
28. West D. *Introduction to Graph Theory*. Prentice Hall; 2017.
29. Dijkstra EW. A note on two problems in connexion with graphs. *Numer Math*. 1959;1:269-271. 10.1007/bf01386390
30. Eagle N, Macy M, Claxton R. Network diversity and economic development. *Science*. 2010;328:1029-1031.
31. Joosten RP, te Beek TAH, Krieger E, et al. A series of PDB related databases for everyday needs. *Nucleic Acids Res*. 2011;39:D411-D419.
32. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983;22:2577-2637.
33. Morcos F, Pagnani A, Lunt B, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A*. 2011;108:E1293-E1301.
34. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *Elife*. 2014;3:e02030.
35. Kryshtafovych A, Schwede T, Topf M, et al. Critical assessment of methods of protein structure prediction (CASP)-round XIII. *Proteins*. 2019;87:1011-1020.
36. Kinch LN, Schaeffer RD, Kryshtafovych A, Grishin NV. Target classification in the 14th round of the critical assessment of protein structure prediction (CASP14). *Proteins*. 2021;89(12):1618-1632. https://doi.org/10.1002/prot.26202
37. 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. https://doi.org/10.1002/prot.26172
38. 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. https://doi.org/10.1002/prot.26171

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Ruiz-Serra V, Pontes C, Milanetti E, Kryshtafovych A, Lepore R, Valencia A. Assessing the accuracy of contact and distance predictions in CASP14. *Proteins*. 2021; 89(12):1888-1900. doi:10.1002/prot.26248