Assessment of three-dimensional RNA structure prediction in CASP15

Rhiju Das[1,2,3,*,‡] https://orcid.org/0000-0001-7497-0972
Rachael C. Kretsch[2,*] https://orcid.org/0000-0002-6935-518X
Adam Simpkin[4,†] https://orcid.org/0000-0003-1883-9376
Thomas Mulvaney[5,6,†] https://orcid.org/0000-0002-4373-6160
Phillip Pham[1] https://orcid.org/0000-0002-0240-6384
Ramya Rangan[2] https://orcid.org/0000-0002-0960-0825
Fan Bu[7,8] https://orcid.org/0009-0000-1903-2270
Ronan Keegan[4,9] https://orcid.org/0000-0002-9495-0431
Maya Topf[5,6] https://orcid.org/0000-0002-8185-1215
Daniel Rigden[4] https://orcid.org/0000-0002-7565-8937
Zhichao Miao[10,11,‡] https://orcid.org/0000-0002-5777-9815
Eric Westhof[12,‡] https://orcid.org/0000-0002-6172-5422

[1]Department of Biochemistry and [2]Biophysics Program, Stanford University School of Medicine, CA USA, [3]Howard Hughes Medical Institute, Stanford University, CA USA, [4]Institute of Systems, Molecular & Integrative Biology, The University of Liverpool, UK, [5]Centre for Structural Systems Biology (CSSB), Leibniz-Institut für Virologie (LIV) and [6]University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany, [7]Guangzhou Laboratory, Guangzhou International Bio Island, Guangzhou 510005, China, [8]Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230036, Anhui, China, [9]Life Science, Diamond Light Source, Harwell Science, UK, [10]GMU-GIBH Joint School of Life Sciences and [11]The Guangdong-Hong Kong-Macau Joint Laboratory for Cell Fate Regulation and Diseases, Guangzhou Laboratory, Guangzhou Medical University, and [12]Architecture et Réactivité de l'ARN, Institut de Biologie Moléculaire et Cellulaire du CNRS, Université de Strasbourg, F-67084, Strasbourg, France. [*]Equally contributing authors. [†]Equally contributing authors. [‡]Correspondence to: rhiju@stanford.edu, miao_zhichao@gzlab.ac.cn, and eric.westhof@ibmc-cnrs.unistra.fr.

Submission to the CASP15 special issue of *Proteins*

Abstract

The prediction of RNA three-dimensional structures remains an unsolved problem. Here, we report double-blind assessments of RNA structure predictions in CASP15, the first CASP exercise in which RNA modeling was assessed. Forty two predictor groups submitted models for at least one of twelve RNA-containing targets. These models were evaluated by the RNA-Puzzles organizers and, separately, by a CASP-recruited team using metrics (GDT, lDDT) and approaches (Z-score rankings) initially developed for assessment of proteins and generalized here for RNA assessment. The two assessments independently ranked the same predictor groups as first (AIchemy_RNA2), second (Chen), and third (RNAPolis and GeneSilico, tied); predictions from deep learning approaches were significantly worse than these top ranked groups, who did not use deep learning. Further analyses based on direct comparison of predicted models to cryogenic electron microscopy (cryo-EM) maps and X-ray diffraction data support these rankings. With the exception of two RNA-protein complexes, models submitted by CASP15 groups correctly predicted the global topology of the RNA targets. Comparisons of CASP15 submissions to designed RNA nanostructures as well as molecular replacement trials highlight the potential utility of current RNA modeling approaches for RNA nanotechnology and structural biology, respectively. Nevertheless, challenges remain in modeling fine details such as non-canonical pairs, in ranking among submitted models, and in prediction of multiple structures resolved by cryo-EM or crystallography.

1 Introduction

Soon after the establishment of the cloverleaf structure of transfer RNA[1,2], three-dimensional models of RNA structures appeared[3,4]. However, it took more than ten years before the first refined experimental structures of the 76-nt yeast $tRNA^{Phe}$ were published [5,6]. For many years, X-ray crystallographic structures of RNA nucleosides and nucleotides allowed us to grasp the fundamentals of RNA stereochemistry. After 1995, following progress in chemistry and X-ray technology, a steady stream of RNA structures with sizes equivalent or larger than tRNAs, culminating with fully functional ribosome structures, revealed the many intricacies of RNA architectures. In parallel, computer programs for RNA modeling appeared (for overview, see ref.[7]). However, it was not until 2011 that a systematic assessment of models prior to publications of the X-ray structures, called RNA-Puzzles, was set up[7,8]. Since not enough targets were available for a short CASP-like season, the Puzzles were organized to occur right as the structures were solved. Since then, several additional publications have reported the results of the RNA-Puzzles assessments[9,10,11]. In 2021, it became clear that accelerations in RNA structure determination[12] would allow enough targets for a single CASP season. Here we report on the first collaborative effort between CASP and RNA-Puzzles teams on a set of RNA targets. Following the success of AI-based tools in protein structure prediction[13] and a surge of interest in RNA during the COVID pandemic[14], the hope of the organizers and assessors was to generate motivation and attention from protein modeling groups to develop and evaluate methods for RNA.

Between April and July of 2022, sequences of twelve RNA targets were received from experimental contributors and disseminated on the CASP website. Models were submitted by over 40 groups, and a double-blind assessment was carried out. Inspired by prior joint

assessments by CAPRI and CASP for protein complexes (see, e.g. refs.[15–18]), two assessments were carried out for RNA: one assessment was performed by the RNA-Puzzles team (Z. Miao & E. Westhof) and a completely independent analysis was performed by assessors nominated by the CASP organizers (R. Das and team). During a dedicated assessors' meeting in October 2022, the two assessments' results were critically compared and this comparison revealed a striking consensus in rankings and choice of top predictors, despite the use of distinct metrics and ranking schemes. Further analysis based on visual inspection of RNA-protein targets, direct comparison to cryogenic electron microscopy (cryo-EM) maps, and molecular replacement trials for targets solved by X-ray diffraction – catalyzed by the general CASP15 conference in December 2022 – revealed additional insights into the limitations and potential of current RNA 3D modeling, which are described here. The identification of accurate models also led to insights by CASP15 RNA experimental contributors and development of novel methods for cryo-EM model refinement, described in two separate papers co-submitted to the CASP15 special issue.

2 Methods

2.1 Computation of RNA-Puzzles-style metrics

The RNA-puzzles-style assessment relied mainly on the Root Mean Square Deviation (RMSD) measure complemented by the Deformation Index (DI)[19]. The RMSD is the usual measure of distance between all atoms (excluding H atoms) of the two superimposed structures. The DI score complements the RMSD values by introducing features specific to RNA in the metric in the following way. The pairs formed by the nucleotides are identified, counted, and annotated in the experimental structure. They are broadly classified as either of the Watson-Crick complementary type (WC, comprising AU, GC, or GU pairs whose geometry are compatible with the standard Watson-Crick-Franklin double helix) or of the non-Watson-Crick type (NWC). The base–base network, i.e. WC, NWC, and stacking interactions in both solution and predicted models are extracted using the MC-Annotate[20] tool. We then compute, for each of the three types of base-base interactions, the number of correctly predicted pairs, the true positive (TP), the number of predicted pairs with no correspondence in the solution model, the false positive (FP), and the number of pairs in the solution model that are not present in the predicted model, the false negative (FN). The Interaction Network Fidelity (INF) is then computed as the Matthews Correlation Coefficient, the geometric mean of the positive predictive value and sensitivity as in Gorodkin[21,22]:

$$INF = \sqrt{\left(\frac{TP}{TP+FP}\right) \times \left(\frac{TP}{TP+FN}\right)},$$

The Deformation Index is then computed as: RMSD/INF. Several partial INF values (and respective DI) can be computed considering only the Watson-Crick (WC) base pairs (INF_{WC}), the non-Watson-Crick (NWC) base pairs (INF_{NWC}), both WC and NWC base pairs (INF_{BPS}), or the stacking interactions (INF_{STACK}). Finally, the Deformation Profile is a distance matrix

computed as the average RMSD between the individual bases of the predicted and the reference models while superimposing each nucleotide of the predicted model over the corresponding nucleotide of the reference model one at a time. It is computed using the ''dp.py'' command from the ''SIMINDEX'' package[19]. For simplification, we also calculate the sum, mean and median of the deformation profile to account for the general accuracy of the prediction. The stereochemical correctness of the predicted models was evaluated with MolProbity[23], which provides quality validation for 3D structures of nucleic acids. MolProbity performs several automatic analyses, from checking the lengths of H-bonds present in the model to validating the compliance with the rotameric nature of the RNA backbone[23,24]. As a single measure of stereochemical correctness, we chose the clash score, i.e., the number of all types of steric clashes per thousand residues[25]. The assessment also considered the coordinate comparison metric TM-score as computed in RNA-Align[26] and the Mean of Circular Quantities[27] to assess accuracy in the torsion angle space. All the source codes and an example notebook are available at: https://github.com/RNA-Puzzles/RNA_assessment.

2.2 Computation of CASP-style metrics

Independently from the RNA-Puzzles-style computations, we assessed the accuracy of the submitted models in a manner closer to recent CASP assessments for protein structure prediction through Z_{RNA}, a weighted Z-score average of several different assessment metrics. To perform the Z_{RNA} evaluation, we develop the casp-rna (CASP tools for RNA) pipeline, which encompasses our workflow for data wrangling, job parallelization, and ranking visualizations. In consideration of RNA as a flexible molecule in which irregular loops may affect RMSD measures, Z_{RNA} explored additional metrics beyond RMSD to capture the topology, local environment, and geometries of RNA. We selected the following tools for our ranking schemes: (1) US-align[28], which was used to compute TM-score through a heuristic alignment approach improving on the original RNA-align[26]; (2) Local-Global Alignment[29] which yielded GDT-TS, the average percentage of aligned C4' atoms (rather than that of C□ in proteins) at cutoffs of 1 Å, 2 Å, 4 Å, and 8 Å; (3) rna-tools[30], a toolkit used to determine the accuracy of contact classifications among base stackings, Watson-Crick interactions, and non-canonical interactions. INF scores were calculated from interaction predictions dependent on ClaRNA[31]; (4) OpenStructure[32,33], a framework used to find lDDT, a metric that measures structural similarity (unlike for proteins, our implementation of lDDT for RNA did not penalize for stereochemical violations); and (5) PHENIX, which reports a clashscore metric for all non-hydrogen bonds overlap worse than 0.4 Å[23,34]. INF scores were computed with ClaRNA to help increase robustness of base pair assignment for low resolution models; these values were slightly different than but highly correlated with INF scores computed with MC-Annotate, the tool normally used by RNA-Puzzles (**Supplemental Figure 1**).

Similar to the assessment of protein models in past CASP assessments, we employ a two-pass procedure for Z-scores[13,35]. For each target and for each of the five metrics, the Z-score (difference with the mean, normalized by the standard deviation) was calculated by taking the mean and standard deviation for the best model from every group. To prevent distortion from very poor outlier predictions, models with initial Z-scores that fall under a tolerance threshold of -2 were discarded. In the second pass, models from the reduced dataset with $Z < -2$ are

assigned Z = -2. To prevent penalization of novel methods that might give poor models for some targets, the sums of just the positive Z-scores of all targets were used to make final rankings. For targets where experimentalists provided multiple conformations to either represent experimental uncertainty or bona fide conformational diversity (e.g., different copies in the crystallographic asymmetric unit or multiple conformations captured by cryo-EM[36]), predictor models were compared to all available experimental models. Groups were rewarded based on their best score. Code for the analysis of submitted models, assessment tools, and documentation using casp-rna (the CASP-style assessment pipeline for RNA) are available as an open-source repository at https://github.com/DasLab/casp-rna. Metrics are also available for interactive viewing on the CASP15 website at https://predictioncenter.org/casp15/results.cgi?tr_type=rna.

2.3 Computation of map-to-model metrics for cryo-EM targets

All models for the 6 targets determined by cryogenic electron microscopy (cryo-EM) (R1126, R1128, R1136, R1138, R1149, and R1156) were assessed directly against the cryo-EM maps. The RNA-protein targets (R1189 and R1190) were excluded from this analysis because no models fit sufficiently well into the density to give robust alignments, but in principle this analysis is compatible with RNA-protein targets. First, models were fit into maps using two approaches. Models were aligned to the target model using US-align[28] and then fit locally using the command *fitmap* in ChimeraX[37]. We also tested an iterative phenix.dock_in_map[38] procedure. For the well-fitting models, there was very little difference between these two methods and thus the *fitmap* method was selected. The following programs were used to measure the listed metrics, in all cases using default parameters (1) Phenix[38], for cross-correlation of the map and model masked by the area around the model (CC_{mask}), cross-correlation of the N highest density peaks in the model-generated map to the map (CC_{volume}), cross-correlation of the N highest density peaks in the model-generated map and N highest density peaks in the map (CC_{peaks}), and model-to-map Fourier shell correlation (FSC) values; (2) TEMPy[39], for cross-correlation coefficient (CCC), mutual information (MI), least-square fit (LSF), envelope score (ENV), and segment-based Mander's overlap coefficient (SMOC); (3) ChimeraX[37] and in-house script. for atomic inclusion[40] and density occupancy; and (4) MapQ[41], for Q-score. An RMSD filter was selected for each target based on visual inspection. Ranking of all the models was carried out by Z-score, following the two-pass procedure described in section 2.2. Code for the analysis can be found at https://github.com/DasLab/CASP15_RNA_EM.

2.4 Scoring against X-ray data and molecular replacement (MR)

All models for the four targets determined by X-ray crystallography (R1107, R1108, R1116 and R1117) were assessed directly against the X-ray data by superimposing them on the target structure with RNAalign[26] and calculating the Log Likelihood Gain (LLG) with respect to the diffraction data using Phaser[42]. For R1108 and R1117, with two RNA molecules in the asymmetric unit, the LLG was calculated for a single copy of the model ideally placed on chain A. A ranking of groups was derived from Z-scores computed from equal weighting of LLG, TFZ (translation-function Z-score from the model search), and CC (correlation coefficient of the map

based on phases from the ideally placed model compared to the map computed by the experimentalists with their final phases). These ranking Z-scores were based on the same two-pass procedure as described in section 2.2.

Molecular Replacement was carried out using the CCP4 package[43] via CCP4 Cloud[44] and specifically the programs Phaser[42] and Molrep[45]. Map correlation coefficients were calculated with the phenix.get_cc_mtz_pdb tool[38]. For R1117, the models were used unedited from all groups. For the other targets, where overall modeling was less accurate, different editing approaches were used with the models from group TS232 (AIchemy_RNA2). For R1107 and R1108, RNA model superposition was carried out with Theseus[46] and nucleotides with higher structural variance values were removed in 10% intervals. The group 232 model_1 after removal of 10, 20, 30, 40 or 50% of nucleotides with highest structural divergence across the models was then used as a search model. MR also made use of models of the U1 small nuclear ribonucleoprotein A protein (U1ABD) component, which were generated using the AlphaFold 2 [47] network in its local ColabFold implementation[48]. For R1116, a version of Slice'N'Dice [49] modified to work with RNA inputs was used to split model 1 from group TS232 into three structural segments using the Birch algorithm from the SciKit toolbox[50].

3 Results

3.1 Classification of the difficulties and qualities of the targets

In **Table 1**, the twelve targets are gathered along with notes on protein and ligand binding, evidence for multiple conformations, and experimental technique and resolution. The difficulty was considered as "easy" when homologous structures were present in the PDB and as of "medium" difficulty when the similarity was mainly due to the function (e.g., the CPEB3 ribozymes self-cleave like a ribozyme of known structure from hepatitis delta virus). Two targets were ranked as "difficult" since no homologous structures had been published and the number of nucleotides was larger than 120. Finally, a fourth "non-natural" category was considered for targets that were human-designed and not found in nature, since it was not clear *a priori* whether these cases would be easy or difficult to model. The majority of targets (8) were solved by cryo-EM, with the rest (4) by X-ray crystallography.

Name	Type	Length	Stoichiometry	Experimental method[a] (resol., Å)	Clash score, expt.	RNA type	Multiple conformations?	Difficulty	Best RMSD (Å)	Notable Groups
R1107	RNA[b]	69	A2	X (2.8)	5	CPEB3 ribozyme (human)	(small differences with R1108)	Medium[c]	4.5	TS232
R1108	RNA[b]	69	A2	X (2.2)	2.4	CPEB3 ribozyme (chimpanzee)	Two conformations in asymmetric unit	Medium[c]	4.5	TS232
R1116	RNA	157	A1	X (3.0)	20.1	Cloverleaf RNA		Difficult	4.8	TS285
R1117	RNA / Ligand[e]	30	A1	X (2.3)	4.2	$PreQ_1$ class I type III riboswitch		Easy[d]	2.0	TS287
R1126	RNA[e]	363	A1	E (5.6)	70.4	Traptamer		Non natural	8.9	TS232
R1128	RNA	238	A1	E (5.3)	1.8	Paranemic Crossover Triangle		Non natural	4.3	TS232
R1136	RNA[e]	374	A1	E (4.5)	0	Apta-FRET	RNA with ligand bound and without.	Non natural	7.2	TS232
R1138	RNA	720	A1	E (5.2)	0.1[f]	6HBC	Kinetically-trapped young and mature	Non natural	7.8	TS232
R1149	RNA	124	A1	E (4.7)	0.25	SARS-CoV-2 SL5	10 models capturing modeling uncertainty.	Difficult	6.9	TS110
R1156	RNA	135	A1	E (5.8)	0.5	BtCoV-HKU5 SL5	4 maps capturing flexibility; 10 models per map capturing modeling uncertainty.	Difficult	5.4	TS128
R1189	RNA/ protein	173	A1B6	E (3.8)	2.6	RsmZ-RsmA	Small differences in R1189/R1190 RNA	Difficult	16.6	TS229[g]
R1190	RNA/ protein	173	A1B4	E (4.6)	1.8	RsmZ-RsmA	Small differences in R1189/R1190 RNA	Difficult	16.0	TS229[g]

[a] X = X-ray crystallography, E = cryo-electron microscopy.

[b] Constructs for R1107 and R1108 both included engineered loops to complex with U1A protein, which aided crystallization; these were not noted as RNA/protein targets during the prediction season.

[c] Known to be related to HDV ribozyme[51].

[d] Three types of $PreQ_1$ riboswitches are known in class I: high resolution X-ray structures of types I & II were known (PDB: 7REX[52], 3Q50[53])

[e] R1117, R1126 and R1146 were noted as RNA/ligand targets during prediction season. A K^+ ion in a G-quadruplex in R1126 and small molecules bound to aptamers displayed in R1136 were not well-resolved in their respective cryo-EM maps and not assessed. Assessment of the pre queuosine ligand in R1117 is included in the overall CASP15 assessment of ligand binding, described separately (Patrick Walters, in preparation).

[f] The model of the mature conformation has a clashscore of 0.09 and the top 10 CASP15 predictions matched this model better than the early conformation (clashscore 63.7).

[g] Similar RMSD predictions came from TS229, TS239 and TS439, all submitted by the same group (Yang).

TABLE 1. Summary and descriptions of the 12 RNA targets in CASP15.

3.2 Assessment and ranking based on RNA-Puzzles metrics

The RNA-Puzzles assessment recognizes that RNA architecture results from a set of coherent interaction networks stabilizing a given fold. There are several interaction networks: the set formed by all Watson-Crick pairs, the set of contacts formed by the stacking between the bases, and finally the set formed by the non-Watson-Crick pairs, the interactions characteristic of tertiary folding. In a 3D structure, the set of Watson-Crick is not always the one predicted because in the folded structure, pairs at the extremities of the helical segments can either disappear or new ones can be formed. The correct choice of stacking between nucleotides or helices is critical for the overall global fold of the RNA. A wrong choice in the helices of the core can lead to very different folds from the native one. Finally, the appropriate positions and orientations of several elements allow for specific non-Watson-Crick pairs to form and lock in the native structure. An approximate association of helices may yield a molecular shape or envelope roughly similar to the native structure, but generally more open and much less compact than the native fold. In such cases, the key sequence conservations that maintain the actual native RNA fold are neither observed nor understood from the modeled structure. Therefore, in addition to using RMSD as a major metric for assessment, the analyses also included distinct metrics that are more sensitive to the interaction networks that comprise RNA.

Table 1 gives the best RMSDs reached by the modeling groups for the 12 targets; they range between 2 Å and close to 17 Å, with many models being in the range between 4.3 Å and 8.3 Å. The trend follows the difficulty level of the targets. Interestingly, for the non-natural designed RNAs, the RMSDs reached are below 8.3 Å. It can be recalled that in a double stranded RNA helix, the average distance between two successive phosphate groups is 7 Å. However, broadly speaking, except for targets R1189 and R1190 (for which the RMSDs reached are beyond 16 Å, see **Table 1**), the overall folding shapes are reproduced, as can be seen in **Figure 1** where all targets are superimposed on the best predicted model as ranked by RMSD.

Supplemental Table 1 presents the number of times that each of the modeling groups produced the 1st, 2nd, or 3rd best model as scored by the various metrics. Separate analyses are shown, based on the best of all five models from each predictor group and based solely on each groups' model 1. Taking a weighted sum of these placements (with weights of 3, 2, and 1 assigned for placing 1st, 2nd, or 3rd) enables ranking of the groups. Whatever the way of counting or of scoring, even with methods that used metrics besides RMSD, two groups consistently reached the first and second ranks, TS232 (AIchemyRNA_2) and TS287 (Chen), respectively. The groups TS081 (RNApolis) and TS128 (GeneSilico) appear both at third positions. Considering those predictions with best RMSD that were ranked first amongst a set of all models submitted (up to five from each group), the groups TS232, TS287, TS081, and TS128 are the top four, with the other groups having weighted sums 50% lower. Among the latter, considering only at best RMSD rankings, TS229 (Yang-Server), TS416 (AIchemy_RNA), TS239 (Yang-Multimer), and TS439 (Yang) occupy the middle range.

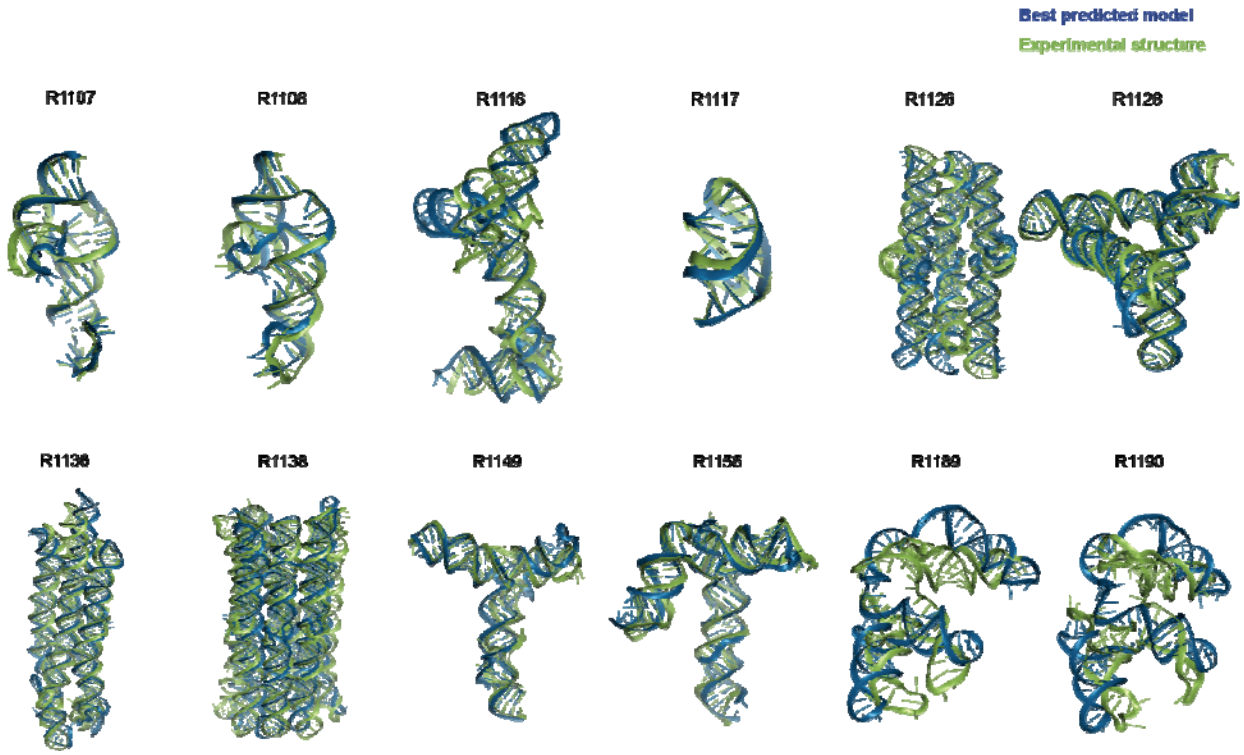

FIGURE 1. Display of all CASP15 RNA targets (green) with the best-ranked model (blue) superimposed for each, chosen based on RMSD comparison of all five predicted models from all predictor groups compared to all available experimental structures. For ease of visualization of RNA topology, protein binding and small molecule ligands (see **Table 1**) are not shown.

3.3 Assessment based on CASP-style metrics

Over 14 prior CASP events, the protein structure modeling community has developed tools to robustly assess predicted models compared to experimental structures. In a second assessment fully independent of the assessment based on RNA-Puzzles above, we explored use of these CASP-style metrics, expanded in use from protein to RNA. For evaluating the topology (*i.e.,* global fold) of predicted RNA structures, we computed the template modeling score (TM-score[26,28]) and the global distance test (GDT[29]). For the latter, we focused on the GDT score for tertiary structure (GDT_TS) rather than the high-accuracy GDT score (GDT_HA[54]) since the RNA models lacked nucleotide-level, much less atomic accuracy. To evaluate models' local quality, to complement the RNA-specific INF score described in section 3.2, we used the Local Distance Difference Test (lDDT[32]) score, which compares distances between nearby atoms in the predicted structure versus the experimental structure and may generalize well between proteins and nucleic acids.

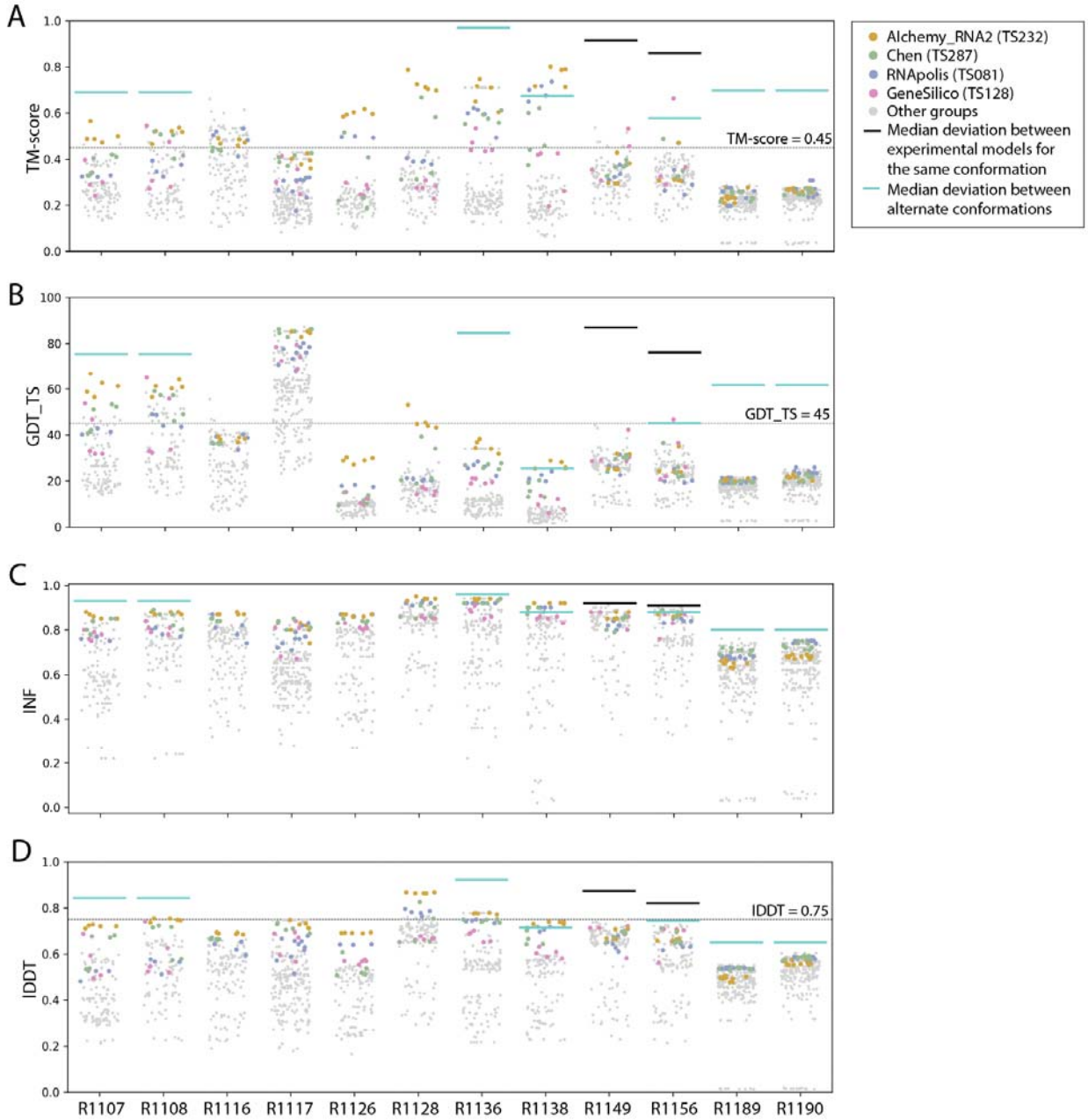

FIGURE 2. TM-score, GDT_TS, INF, and lDDT values for all targets. Scores for all models submitted for all targets are depicted. Models from the four top performing groups are highlighted as colored points, and all other groups' models are shown as gray points. Cyan lines indicate the median deviation between experimentally determined models for alternate conformations (see main text), and black lines indicate the deviation between alternate models derived from experimental data for the same conformation.

The topology metrics (TM-score and GDT_TS) suggest that all targets, aside from the two RNA-protein complexes R1189 and R1190, elicited some predicted models that recovered correct global folds, based on criteria that have been previously discussed in the context of RNA template identification (TM-score > 0.45[26,28], **Figure 2A**) or protein topology assessment (GDT_TS > 45[55], **Figure 2B**). We note that these criteria for 'correct fold' may not apply at the extremes of lengths for our RNA targets. On one hand, the "easy" $PreQ_1$ riboswitch target

(R1117) is small with only 30 nucleotides, and the TM-score values, which involve a length-dependent distance parameter, are much lower than GDT_TS values (**Figure 2A,B).** The accuracies reflected by GDT_TS match expected accuracies gauged by visual examination. On the other hand, models that visually captured correct folds for large designed RNA's (R1126, R1128, R1136, R1138) were properly assigned high TM-scores, while GDT_TS scores were mostly lower than 45 (**Figure 2A-B**). For predictor models for a given target, the TM-score and GDT_TS correlated well, but the relationship between the two varied across different targets (**Figure 3A**). These comparisons suggest that TM-score and GDT are useful for ranking models for a given target but thresholds for 'good' TM-score and GDT may need recalibration for very small and very large RNA molecules, respectively.

As an alternative metric for model quality that might generalize between protein and RNA, we considered lDDT. While not directly measuring global topology, lDDT has been used as a primary accuracy indicator in numerous prediction contexts, including CAMEO, where a threshold of lDDT > 0.75 is used to denote a good match when comparing templates to target structures and to assign difficulty.[56,57] Across all targets, lDDT values for best predictions ranged from 0.5 to 0.9, again with the lowest performance in RNA-protein complexes (**Figure 2D**). Interestingly, for the 10 RNA-only targets, CASP15 predictors achieved models with lDDT close to 0.75, and visually excellent models for the small, "easy" target R1117, the "medium" target R1108, and the "non-natural" and larger targets (R1128, R1136) achieved the 0.75 threshold. For future CASP, CAMEO, and other modeling challenges, lDDT may provide the most cleanly interpretable measure of accuracy, with a cutoff of 0.75 applicable across nucleic acids and proteins.

These CASP-inspired metrics correlated well with RNA-puzzle based metrics described in section 3.2. For topological metrics, while RMSD and GDT_TS are not linearly correlated (**Figure 3A**), they have positive rank-based correlation (Spearman correlation coefficient 0.61, **Figure 3B**). The local interaction metrics, INF and lDDT, correlate excellently (Spearman correlation coefficient 0.91, **Figure 3B**) in what seems to be a near-linear and size-independent relationship (**Figure 3A**). This is a remarkably strong correlation; INF focuses on a selection of RNA-specific interactions while lDDT compares all heavy-atom distances for atom pairs within 15 Å, a similar length scale to the distances across base pairs monitored by INF. This observation suggests that lDDT may capture the subset of interactions measured in INF while allowing generalization across protein and nucleic acids. Finally, if we compare topological metrics with local metrics, we still maintain a positive correlation (Spearman correlation coefficient 0.67-0.87, **Figure 3B**), however the relationship is non-linear; local metrics like lDDT are able to discriminate models with low accuracy while topological metrics like GDT_TS are better able to discriminate the high accuracy models (**Figure 3A**).

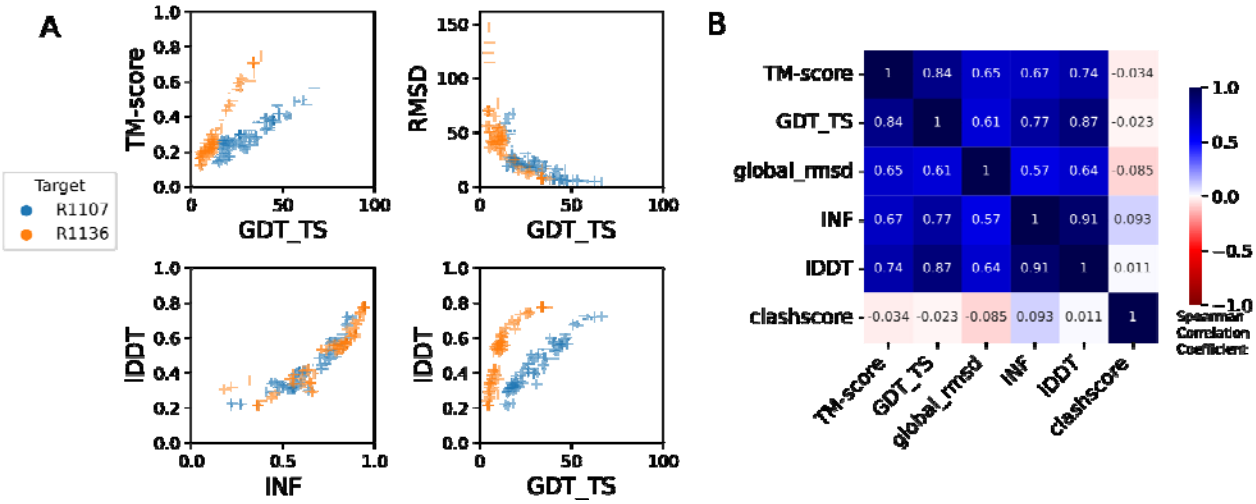

FIGURE 3: Comparison of assessment metrics for RNA targets. (A) Scores for all models for representative targets R1107 (blue) and R1136 (orange): top-left GDT_TS vs. TM-score, top-right GDT_TS vs. RMSD, to compare across topology-level metrics; bottom-left INF vs. lDDT compares the two local metrics; and bottom-right GDT_TS vs. lDDT compares topology-level to local metrics. (B) Average Spearman rank correlation coefficient (calculated separately per target, then averaged over each target) between each pair of scores labeled on each row and column, colored by high positive correlation (dark blue), no correlation (white), or negative correlation (red).

To provide a more quantitative threshold for good model accuracy for each target, we sought to estimate the deviation between experimentally determined structures. Where possible, we measured the deviation in TM-score, GDT_TS, and lDDT between distinct experimentally captured conformations (cyan lines in **Figure 2**). More specifically, we compared the following structure pairs in targets with multiple conformations (see also **Table 1**): the point-mutations for the CPEB3 ribozyme[58] (R1107 vs R1108), the apo and holo structures of the aptamer Apta-FRET[59] (R1136), the young and mature conformations of 6HBC[60] (R1138), the four cryo-EM classes for the SL5 domain of the bat coronavirus HKU5 (R1156), and finally the RNA structures for the RsmZ-RsmA RNA-protein complexes with six vs. four proteins bound (R1189 vs R1190). In addition, for two cases, we measured the deviation between different models derived from the same experimental data (black lines in **Figure 2**), comparing distinct models built into cryo-EM density maps for the SL5 domains of SARS-CoV-2 (R1149) and BtCov-HKU5 (R1156). In some cases (R1138, R1156), predicted models reached TM-score, GDT-TS, and lDDT values comparable to the deviation between distinct experimentally determined conformations (cyan lines, **Figure 2**), though in no case were there models whose accuracies exceeded the experimental precision expected for a single captured conformation (black lines, **Figure 2**).

To rank the performance of predictors, we developed a Z-score metric that enabled combined evaluation of models' topology, local environment, and stereochemical correctness. Our topology scores included the TM-score and GDT-TS, our local environment scores consisted of INF and lDDT, and our stereochemical correctness scores were based on clashscore[25], which

has been used widely for both protein and RNA structural assessment. We used the following weighted sum of scores:

$$Z_{RNA} = \frac{1}{3}[Z_{TM} + Z_{GDT-TS}] + \frac{1}{8}[Z_{INF} + Z_{lDDT}] + \frac{1}{12}Z_{clash}$$

Because we did not expect atomically accurate models in this first RNA round of CASP, we chose to reward models that recover the global fold (high weight for topology terms) compared to those that recover local details (low weight for local environment scores) or produce correct nucleotide geometries (low weight for clashscore). Each group's Z-score for a given target was computed using their best predicted model, and groups' total scores were calculated as the sum of all positive Z-scores across all targets (**Figure 4A**). The top performing predictor groups based on this combined Z-score ranking were AIchemy_RNA2 (TS232), Chen (TS287), RNAPolis (TS081), and Genesilico (TS128). These were the same groups as the top four highlighted by the independent analysis by the RNA-puzzles-style assessment.

Interestingly, the top four groups did not include any server submissions; the top-ranked servers (Ultrafold-server, TS125; and Yang-server, TS229) placed at positions 8 and 9, and gave Z-scores that were less than three-fold lower than the top two predictor groups. Furthermore, based on abstracts collected for the CASP15 conference, while the majority of CASP15 RNA predictors groups tested deep learning methods (orange highlights in **Figure 4A**), the top 4 RNA groups did not use deep learning approaches (see also articles by RNA predictor groups co-submitted for the CASP15 special issue; and https://predictioncenter.org/casp15/doc/presentations/Day3/).

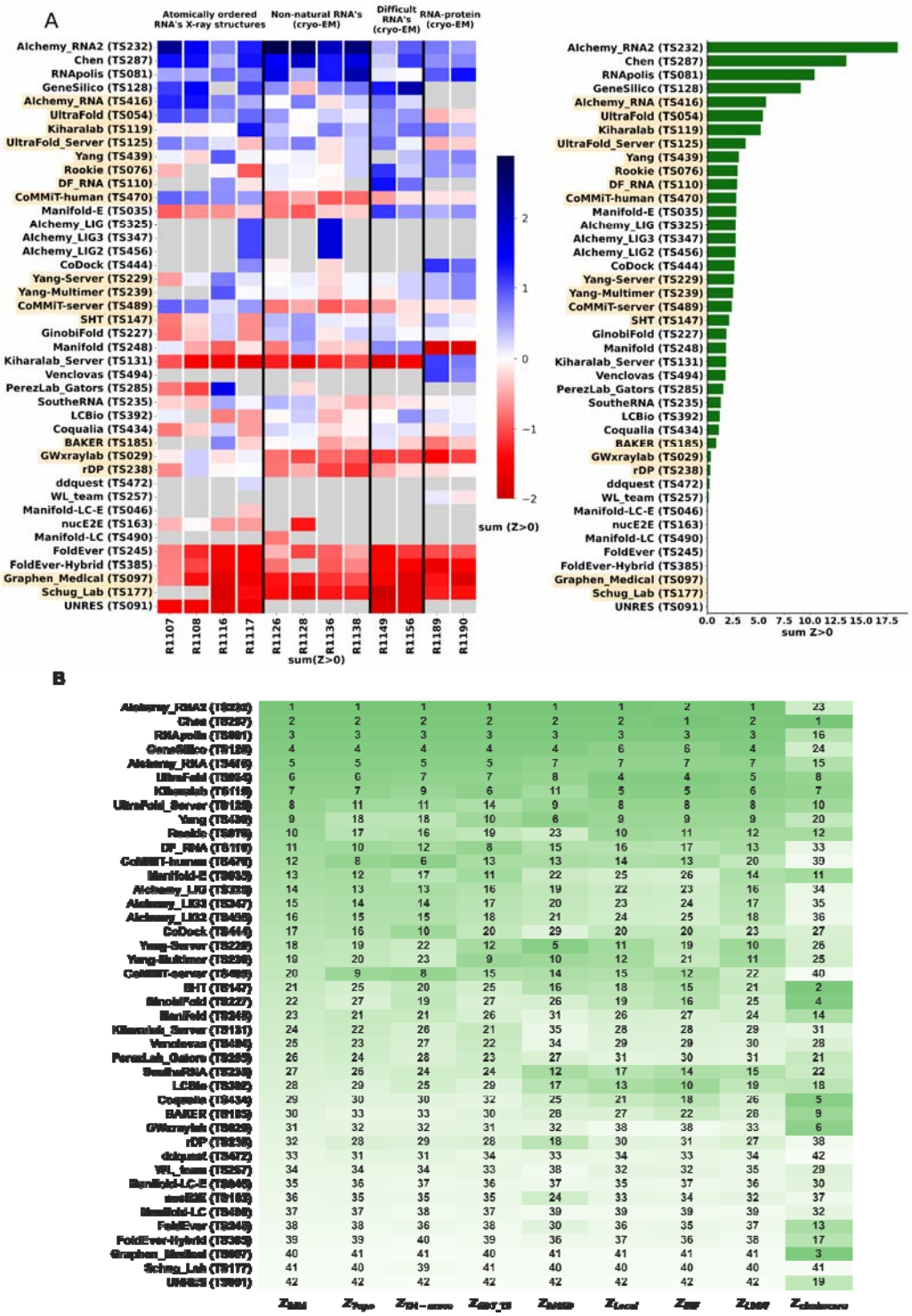

FIGURE 4. CASP-style Z-score based Rankings. (A) Heatmap of groups ranked by Z_{RNA} Groups which used deep learning, as reported in the participant's abstract to CASP15, are indicated in orange. The summation of positive two-pass Z-scores for each of the 12 targets is summarized in the barplot (right). Groups are ordered by their Z_{RNA} rankings. **(B)** Robustness of ranking to different choices in assessment. Columns show group rankings based on subsets of the Z_{RNA} score or individual metrics; coloring reflects rankings under each metric.

Groups	Z_{RNA}	$Z_{TM\text{-}score}$	Z_{GDT_TS}	Z_{RMSD}	Z_{INF}	Z_{lDDT}	$Z_{clashscore}$
AIchemy_RNA2 (TS232)	18.58	22.17	22.89	15.17	12.07	15.12	2.13
Chen (TS287)	13.56	15.49	13.70	13.59	12.46	13.73	6.74
RNApolis (TS81)	10.48	11.83	11.69	10.41	9.80	10.26	3.84
GeneSilico (TS128)	9.14	11.24	10.64	8.45	7.53	8.74	2.06
AIchemy_RNA (TS416)	5.73	6.25	5.72	6.36	6.55	6.21	4.07
UltraFold (TS54)	5.45	4.62	5.52	6.33	9.34	8.58	5.46
Kiharalab (TS119)	5.21	3.83	5.61	5.64	9.29	7.66	6.11
UltraFold_Server (TS125)	3.73	3.44	3.41	6.04	6.07	5.38	5.21
CoMMiT-human (TS470)	3.09	2.50	3.78	6.50	5.36	5.31	3.12
DF_RNA (TS110)	2.98	2.80	3.19	1.99	4.32	4.18	4.81
Yang-Server (TS229)	2.91	3.30	4.31	3.01	3.41	3.52	0.42
CoMMiT-server (TS489)	2.85	5.36	3.52	3.75	4.28	2.88	0.00
AIchemy_LIG (TS325)	2.83	2.73	3.76	2.12	2.13	3.09	5.04
AIchemy_LIG3 (TS347)	2.77	3.21	3.26	2.35	2.30	2.93	0.16
AIchemy_LIG2 (TS456)	2.77	3.21	3.26	2.35	2.30	2.93	0.16
Yang-Multimer (TS239)	2.77	3.21	3.26	2.35	2.30	2.93	0.16
Yang (TS439)	2.67	3.53	2.80	1.04	3.07	2.57	1.52
Manifold (TS248)	2.62	2.19	3.65	6.92	3.15	5.01	1.54
GinobiFold (TS227)	2.51	2.05	3.98	5.82	3.05	4.62	1.68
SHT (TS147)	2.41	4.34	3.29	3.52	4.28	2.76	0.00
BAKER (TS185)	2.16	2.38	1.98	2.68	3.69	2.77	6.68
Rookie (TS76)	1.86	2.40	1.43	1.61	3.45	2.37	6.66
SoutheRNA (TS235)	1.85	2.24	1.93	0.82	2.09	2.50	4.15
Manifold-E (TS35)	1.84	1.81	2.24	0.36	2.00	1.56	0.60
Coqualia (TS434)	1.78	1.79	2.21	0.37	1.60	1.49	0.73
PerezLab_Gators (TS285)	1.58	1.78	2.06	1.55	1.07	0.98	2.35
LCBio (TS392)	1.34	1.94	1.98	4.10	3.82	2.99	2.19
rDP (TS238)	1.25	1.81	1.21	2.67	4.52	2.90	3.54
GWxraylab (TS29)	1.14	1.53	0.63	1.78	3.18	2.16	6.64
CoDock (TS444)	0.89	0.55	1.10	1.31	2.66	1.60	5.26
ddquest (TS472)	0.36	1.10	0.78	0.42	0.00	0.13	6.11
nucE2E (TS163)	0.30	1.57	1.31	2.67	0.70	2.09	0.00
Manifold-LC-E (TS46)	0.23	1.33	0.12	0.40	0.00	0.02	0.00
Venclovas (TS494)	0.13	0.27	0.13	0.00	0.52	0.00	0.73
WL_team (TS257)	0.00	0.00	0.00	0.28	0.00	0.00	0.71
FoldEver (TS245)	0.00	0.21	0.05	1.95	0.00	0.36	0.00
FoldEver-Hybrid (TS385)	0.00	0.00	0.00	0.00	0.00	0.00	0.48
Kiharalab_Server (TS131)	0.00	0.05	0.00	0.98	0.00	0.00	4.17
Graphen_Medical (TS97)	0.00	0.00	0.00	0.29	0.00	0.00	3.78

UNRES (TS91)	0.00	0.00	0.00	0.00	0.00	0.00	6.68
Manifold-LC (TS490)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Schug_Lab (TS177)	0.00	0.00	0.00	0.00	0.00	0.00	3.54

TABLE 2. Z-scores for predictor groups using different combinations of assessment metrics.

To better understand uncertainties in the rankings, we repeated the Z-score analysis using sub-components of the Z-score. Ranking groups by the two "topology" terms (GDT_TS and TM-score) alone or in combination, or using RMSD, gave rankings with the same top four groups, up to some switching of third and fourth place (**Figure 4B** and **Table 2**). Use of only local environment terms (lDDT and INF) retained the same top three predictor groups, with some groups switching in ranks of the groups after the top three. After the top four, the rankings are less consistent, which is not surprising given the small numerical score differences in these placements (**Figure 4A** and **Table 2**). Ranking groups by clashscore alone did not correlate with the other rankings (**Figure 3B** and **Figure 4B**), presumably because different predictors used somewhat different refinement schemes and were not told *a priori* that they would be assessed on clashscore. Overall, the ranking of the top four groups in CASP RNA structure modeling was robust to changes in metrics used and across two independent assessments.

3.4 Detailed assessment for RNA-protein complexes

The poorer predictions and the presence of RNA-protein contacts for the two RNA-protein complexes R1189 and R1190 largely precluded useful accuracy rankings from the metrics described above, so we carried out a detailed visual assessment for these targets. This assessment involved checking whether predictions had the right nucleotide – amino acid contacts and then visually assessing whether the topology was correct. For the contact-based analysis, a contact was defined as any pair of nucleotide and amino acid containing atoms within 5 Å of one another. The Matthews Correlation Coefficient (MCC) was used to score the contacts made by the predictions against those of the targets. The distribution of scores is shown below. The highest scoring model from each group with MCC scores above 0.1 (roughly the beginning of the non-zero peak in the distribution) were then visually assessed.

For the topology analysis, we needed to establish a well-defined descriptor for the RNA-protein binding arrangement. This was achieved by coloring each protein in the order they are visited – blue (B), then red (R), then green (G) for the three RsmA homodimers in R1189 – and comparing this pattern for each model against the experimental structure (topology string: BRGRGB). In the case of R1190, which involved only two RsmA homodimers, not all six regions of the RNA were bound; in particular, the regions of the RNA at approximately nucleotides 25 and 50 should not interact with a dimer. For R1189, no models exhibited the correct topology for visiting the 6 RsmA proteins (**Table 3**). For R1190 (topology string: B-R-RB), the best model according to the MCC score (MCC = 0.39) predicted the non-interacting RNA regions correctly ('-' in Table 1) but the RNA visited the proteins in the wrong order. Many of the lower scoring models (MCC = 0.21-0.29), did visit the proteins in the correct order but misplaced the non-interacting regions. As judged by this MCC contact-based score supplemented by protein-binding topology analysis, TS119 (Kiharalab) and TS329 (LCBio) produced top-3 models for

both targets (**Table 3**). In contrast, ranking based purely on RNA RMSD highlighted models from TS229 and other models from Yang's group (**Table 1**); these models were less satisfactory from the point of view of protein-RNA contacts, showing the importance of complementary analyses in ranking these very difficult targets.

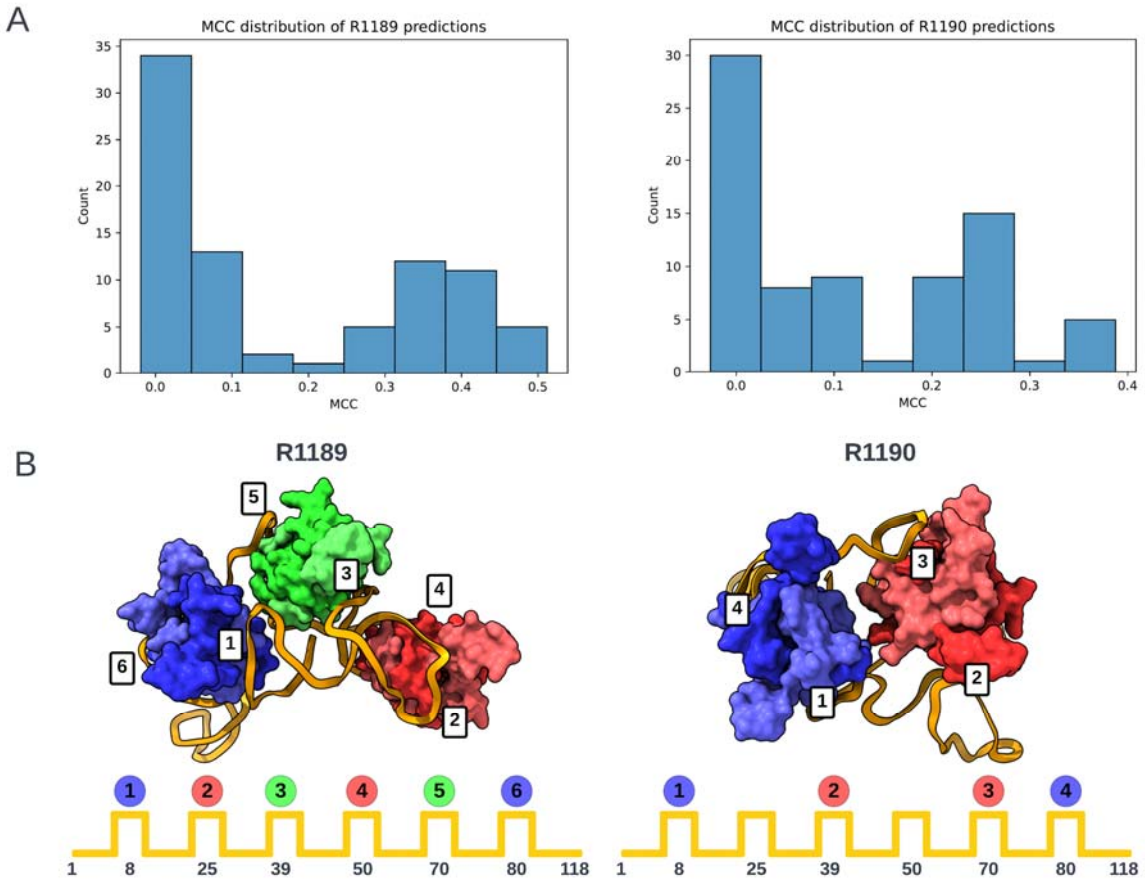

FIGURE 5. (A) A subset of models gave higher Matthews Correlation Coefficients (MCC) for RNA-protein contact accuracy in the two RNA-protein targets R1189 and R1190 (RsmZ-RsmA RNA-protein complexes). (B) Scheme for classifying topology of protein binding to RNA based on order of protein contacts to RNA. Each dimer is assigned a color based on the order it was visited in. Experimental cryo-EM structures are shown at top with positions of binding on RNA diagrammed below.

Target	Group ID	Group Name	Model	Contact MCC	Topology
R1189 (native topology: BRGRGB)	119	Kiharalab	3	0.51	BRGRBG
	232	AIchemy_RNA2	2	0.41	BRGRBG
	392	LCBio	2	0.41	BRGRBG
	444	CoDock	3	0.39	BRGRBG
	439	Yang	2	0.38	BRGRBG
	131	Kiharalab_Server	2	0.37	BRGRBG
	494	Venclovas	4	0.32	BRGRBG
	434	Coqualia	1	0.26	1 hexamer
	185	BAKER	1	0.18	BRGB
R1190 (native topology: B-R-RB)	392	LCBio	5	0.39	B-R-BR
	119	Kiharalab	2	0.29	BR-RB-
	444	CoDock	5	0.27	BR-RB-
	494	Venclovas	4	0.26	BR-RB-
	131	Kiharalab_Server	5	0.26	BR-RB-
	232	AIchemy_RNA2	4	0.21	BR-RB-
	434	Coqualia	1	0.17	Dimers conjoined
	035	Manifold-E	1	0.12	Protein separated from RNA
	227	GinobiFold	1	0.10	Dimers conjoined

TABLE 3. Matthews correlation coefficients and topological assessment of the best model from each group with an MCC greater than 0.1. The symbols B, R, G and '-' indicate blue, red, green and unmodelled regions as per **Figure 5B**.

3.5 Ranking based on direct comparison to cryo-EM maps

The 'native' experimental models built from RNA cryo-EM maps may be particularly susceptible to biases from computational procedures or biases in human interpretation due to the generally low resolution of these maps (see, e.g., experimental model clashscores higher than 10 in **Table 1**, which typically arise from fitting errors). In particular, for RNA, when the cryo-EM map has resolution worse than ~3 Å, the separation between bases cannot be resolved and thus base placement can be highly dependent on the modeling approach used by the experimentalists. We therefore sought to rank CASP predictions based not on comparison to the coordinates provided by the experimenters ('model-to-model') but by comparison directly to the experimental maps ('map-to-model').

For all 6 RNA-only cryo-EM targets, there were models that could visually fit well into the maps (**Supplemental Figure 2**). To determine a quantitative ranking of predictor groups, previously available model-to-map metrics were computed (Methods; **Supplemental Figure 4**). These model-to-map metrics were developed to assess goodness of fit for models prepared with knowledge of maps; many were not designed to account for very poorly fitting models, with unmodeled density and atoms outside density, as we have here. For example, atomic inclusion[40] penalizes predicted atoms that appear outside of density, and correlation coefficient at peaks (CC_{peaks})[34] penalizes density that is not accounted for by a prediction. We attempted to find a combination of scores to balance these problems; however, in the end, we decided that no weighted combination of metrics was sufficient, Although overall correlation of map-to-model metrics to model-to-model metrics was high (**Supplemental Figure 5B**), there were outliers receiving high map scores for poor models by, for example, condensing all atoms into a single small area, most notably group 238 (**Supplemental Figure 3C and 5A**). Thus, as in previous CASP evaluation for cryo-EM of protein targets[61], we used a filter (**Supplemental Figure 3B**), only ranking models that exhibited sufficiently high model-to-model scores. Due to the size dependence of TM and GDT-TS noted above, we decided to set this cutoff based on RMSD. The correlation between metrics was generally improved after this filtering (**Supplemental Figure 3A**).

For ranking, we selected a set of metrics that correlated well with visual inspections of fit and chose the standard measures of cross-correlation, accounting for modeled (CC_{mask}) and unmodeled regions (CC_{peaks}), and scores developed or shown to be most discriminatory for medium-resolution maps, atomic inclusion (AI), mutual information (MI), and Segment based Manders' Overlap Coefficient (SMOC)[62,63]. We note that no metrics tested were RNA specific and can be used to assess any macromolecular complex. We used Z-score-based ranking, previously described, with uniform weight of the selected metrics:

$$Z_{EM} = \frac{1}{5}\left[Z_{CCmask} + Z_{CCpeaks} + Z_{MI} + Z_{SMOC} + Z_{AI}\right]$$

AIchemy_RNA2 (TS232) achieved the highest Z_{EM} score, followed by Chen (TS287), GeneSilico (TS128), and RNApolis (TS081), and then others (**Figure 5A**). This ranking matched with the model-to-model assessment (orange bars in **Figure 5A**). This overall ranking was also maintained, barring group 238, without filtering out poor models (**Supplemental Figure 5A**); however, the filter should be maintained until Z_{EM} is robust to the problematic high scores of condensed models, by for example the inclusion of clashscore.

Overall, the results show that assessing models based on direct comparison to cryo-EM maps, appears feasible and that results are consistent with rankings based on model-to-model comparisons. Direct map-to-model assessments may be particularly important in future CASP events as prediction accuracy increases and approaches the level of detail obtained at typical cryo-EM map resolutions.

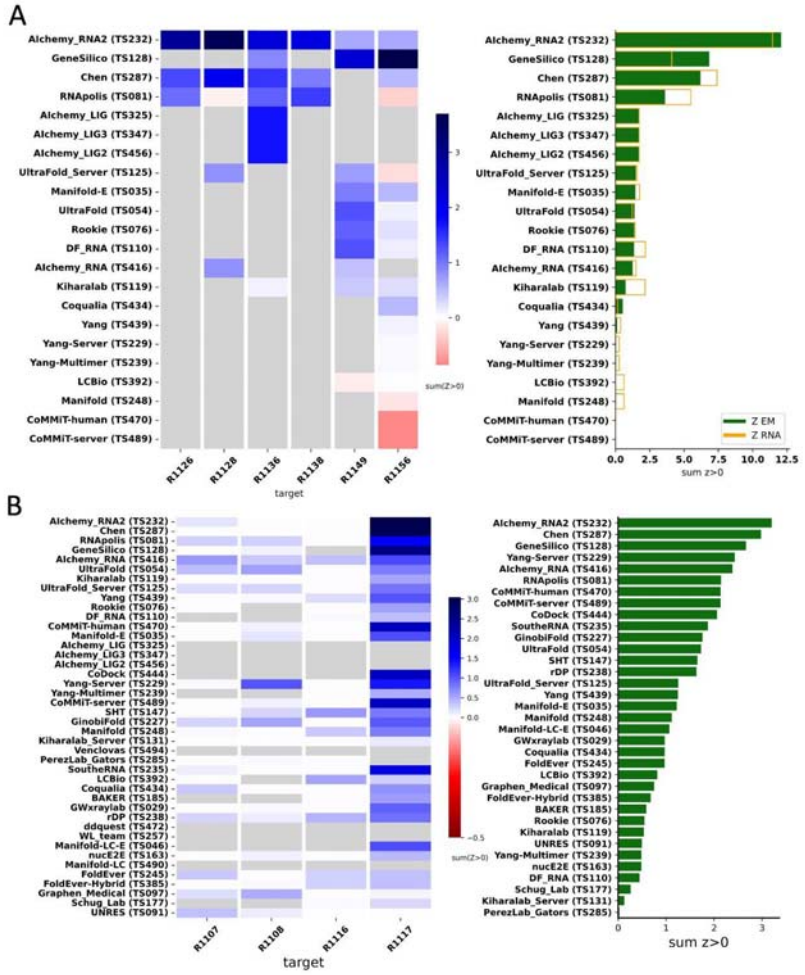

FIGURE 5. Ranking of CASP RNA predictions based on direct comparison to experimental data. **(A)** Ranking of six RNA-only cryo-EM targets based on Z-scores for map-to-model metrics (Z_{EM}). Only a subset of models with clear alignments to maps were included in the comparison; see **Supplemental Figure 5** for analysis over all models. **(B)** Group ranking for X-ray crystal structure targets based on Z-scores for metrics that directly compare the models to the crystallographic data (Z_{MX}).

3.6 Ranking based on direct comparison to crystallographic data

In analogy to the map-based assessment of cryo-EM targets in the previous section, we investigated whether similar comparisons to the experimental data might enable ranking of the four RNA targets solved by X-ray crystallography. Similar to above, the only use of the experimentally derived model was to align predictor models. All predictor models were compared directly to the crystallographic data by first ideally placing the model using RNAalign[26] and then calculating an Log Likelihood Gain (LLG) and translation-function Z-score (TFZ) with Phaser's RNP search[42] and a global map CC with phenix.get_cc_mtz_pdb[38]. We used a Z-based ranking after a round of outlier removal (see Methods) with a uniform weighting of these metrics:

$$Z_{MX} = \frac{1}{3}\left[Z_{LLG} + Z_{TFZ} + Z_{global\ map\ CC}\right]$$

The rankings are most strongly influenced by performance on R1117 since Z_{MX} scores for the other targets were relatively uniform and comparatively poor (**Figure 5B**). The top-ranking groups by this metric were TS232, TS287, and TS128 (AIchemy_RNA2, Chen, and GeneSilico, respectively), which were also the 3 groups that succeeded in follow up molecular replacement trials for R1117; see **Section 3.8** below.

3.7 Topologically accurate CASP15 RNA models miss detailed features and aspects of conformational heterogeneity

Ranking CASP15 RNA predictions based on the quantitative comparisons above highlighted several models for more detailed visual inspection, which revealed the potential and limitations of the models. One example, the chimpanzee CPEB3 ribozyme R1108, illustrates the use of the Deformation Profile and variable accuracy in targets of “medium” difficulty (**Table 1**). In **Figure 6I-J**, the superimposition of the experimental structure with the best model (TS232_4, from AIChemy_RNA2) is shown with the large deviations at the apical loops. The positions of these loops on the Deformation Profile (**Figure 6A-B**) are indicated highlighting the restricted regions with high discrepancies.

One of the highly successful models is that of the paranemic crossover triangle (PTX) R1128, a molecule with no natural homologs whose difficulty for modeling was unclear before the CASP15 results. It is a designed sequence made of four 4-way junctions and a co-axial stack between terminal helices (**Figure 6C-F**). The good achievement can be partly explained by the folding constraints of the design and the use of known structural modules. The helices are regular with known GU pairs and capping UNCG loops, without unpaired or bulging residues (**Figure 6C**). The tight junctions and the bulky RNA helices impose strong constraints on the fold and prevent knot formation (**Figure 6D**). The accuracy of the modeling (TS232_1) with a RMSD of 4.3 Å and an INF of 0.88 is apparent in the deformation profile with a rather uniform deformation throughout (**Figure 6E**). The origins of the main errors are in the twist angles between stacked helices in the 4-way junctions that propagate maximally towards the apical loops (**Figure 6F**). In the experimentally determined structure, at those 4-way junctions, there are H-bonds linking one hydroxyl O2’ atom to an anionic phosphate oxygen of a residue on the crossing strand, maintaining a tight packing. These H-bonds are not present in the modeled structure, leading to a looser packing and slightly larger twist angle (**Figure 6G**). Despite these errors in fine details, the CASP15 blind model TS232_1 was closer to the cryo-EM-derived structure than the original model of the PTX structure designed by Andersen and colleagues (see paper co-submitted to CASP15 special issue by RNA experimentalists).

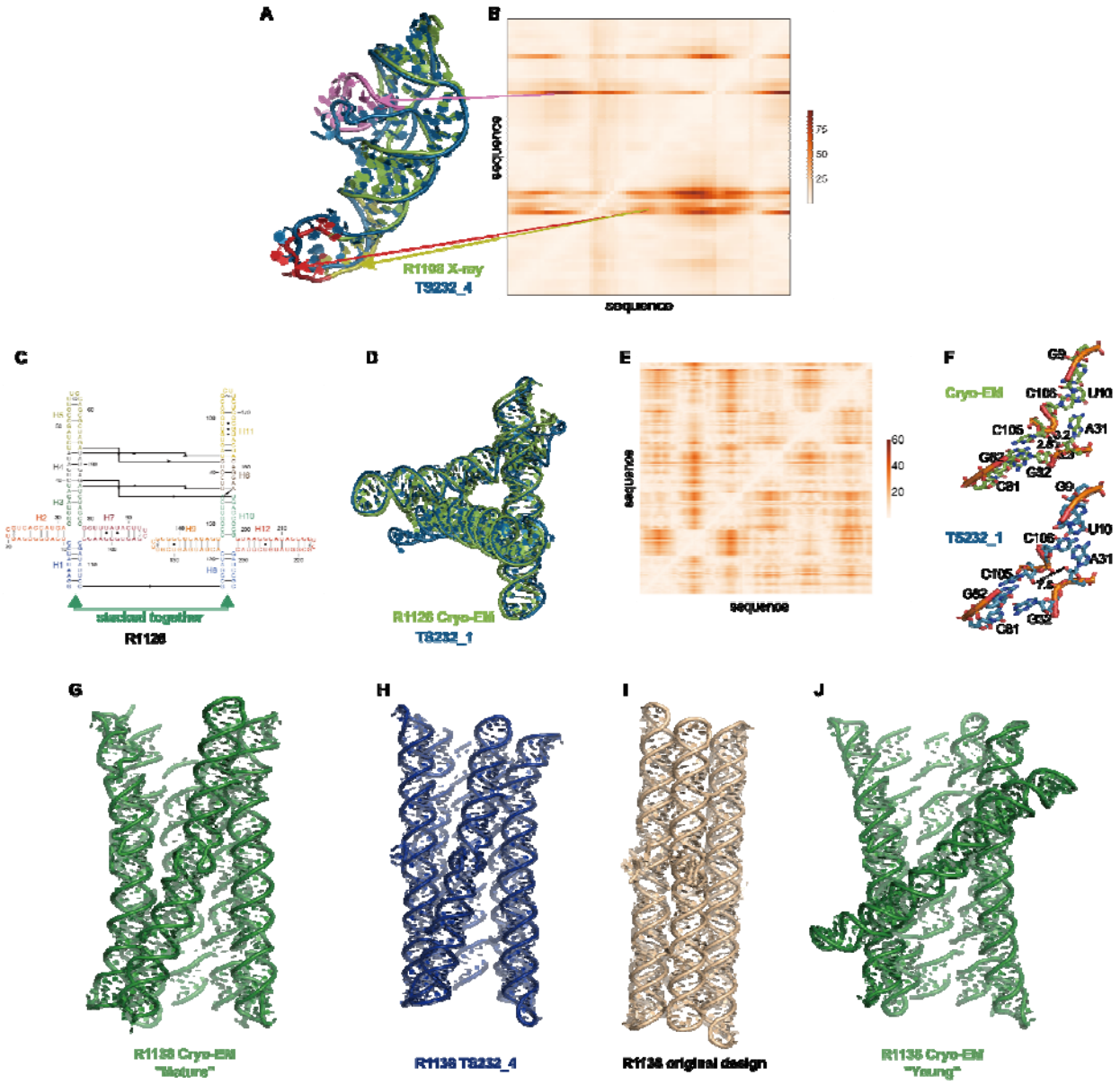

FIGURE 6 Detailed inspection of "medium" and "non-natural" targets. (A) For R1108 (chimpanzee CPEB3 ribozyme), superimposition of the experimental structure (green) with the best model (TS232_4 from AIChemy_RNA2, as blue, RMSD 4.5 Å) is shown. Notice the large deviations at the apical loops (as red, yellow and pink) and their positions on **(B)**, the Deformation Profile. **(C)** Diagram of the secondary structure (2D) of target R1128, a designed paranemic crossover triangle. The helices are numbered from H1 to H11. The secondary structure contains four 4-way junctions. In the two 4-way junctions drawn as "open", helices H1 stack with H2 and H3 with H7 for one 4-way junction and for the second one helices H8 stacks with H9 and H10 with H12. Helices H1 and H8 are stacked together. The pairs between G and U are marked by a dark dot (G•U pair). The Leontis-Westhof[64] symbols are used to annotate the Watson-Crick/Sugar edge pair between G and U in the capping apical 5'UUCG3' tetraloops. **(D)** Experimental structure (green) superimposed on the model TS232_1 (blue) with the lowest RMSD (4.3 Å). **(E)** The deformation profile (see Methods) between the same set of structures (at the right, the color scale where white represents excellent superimposition). The reddish regions indicate where the discrepancies are

largest; they concentrate at the 4-way junctions where the experimental structure is more compact and with H-bonding contacts between the strands than the model structure as shown in **(F). (G-J)** Models for R1128 (Paranemic Crossover Triangle, PXT). Cryo-EM of mature conformation **(G)** agrees better with blind CASP model TS232_4 **(H)** than with original models prepared by this nanostructure's designers **(I)**. Cryo-EM also captured an early folding intermediate **(J)** that was not predicted well by any CASP15 groups.

Indeed, for all four non-natural RNA targets in CASP15 (**Table 1**), the AIchemy_RNA2 group (TS232) submitted models that were visually accurate (**Figure 1**). Furthermore, this group, along with Chen (TS287) and RNAPolis (TS081) were notably separated from other groups, including all automated servers, for these non-natural targets, suggesting that these predictors benefitted from their human intuition to recognize the secondary structures and overall tertiary folds intended by the nanostructures' human designers. Interestingly, in all four cases, the predictor groups were able to blindly predict structures that agreed better with the cryo-EM maps than the original models made by Andersen and colleagues. As another example, for R1138 (six-helix bundle, **Figure 6G-H**), the original design and the cryo-EM structure of the 'mature' form of the RNA agree in overall topology, as reflected by a TM-score of 0.623, well above the 0.45 threshold (**Figure 6G-H)**. Nevertheless, the AIchemy_RNA2 model TS232_4 achieves an even higher TM-score of 0.800 (**Figure 6I**). These results suggest that, despite the lack of natural sequence homologs, "non-natural" RNA targets could be considered "easy" for 3D RNA structure prediction, as long as they are composed of readily identifiable helices and non-canonical motifs.

Interestingly, for the same R1138 six-helix-bundle, cryo-EM also captured a distinct 'young' structure for the RNA (**Figure 6J**) that is dominant immediately after the transcription of the RNA and requires hours to resolve into the 'mature' form. The 'young' and 'mature' structures do not differ in their Watson-Crick-Franklin helices but, to interconvert, would require breaking of a kissing loop interaction, twisting of the two kissing elements about their helical axes, and then reformation of the kissing loop.[60] None of the CASP models produced models close to the 'young' structure. Other natural and designed RNA systems are known to display similar kinetic traps and topological isomers[65,66], and it will be interesting to see if in future CASPs, such conformations can be blindly predicted.

A common theme was that the model ordering as submitted by the predictor groups generally did not correspond to the ranking based on RMSDs (or other metrics) between experimental and model structures. This was the case for the R1108 and R1138 targets noted above, where the fourth models from group TS232, and not the first models, were accurate. The models from group TS110 (DF_RNA) for the "difficult" target R1149 (the SL5 domain from SARS-CoV-2) provides a good additional example. The best RMSD of all CASP15 submissions is model #2 by TS110 as depicted in **Figures 7A-7D**. The RMSD between the experimental structure and TS110_2 is 6.9 Å (superposition shown on **Figure 7A** with the respective Deformation profile on **Figure 7B**). On the other hand, the RMSD between the experimental structure and first model TS110_1 is 21.7 Å. The superposition (**Figure 7C**) and the corresponding Deformation profile (**Figure 7D**) confirm that the global fold of TS110_1 is inaccurate despite its submission as model 1. In particular, the reddish regions indicate where the discrepancies are largest; they

concentrate at the 4-way junctions where the experimental structure is more compact and with H-bonding contacts between the strands than the model structure as shown in Figures **7A** and **7C.** Overall, in 63% of the sets of CASP15 predictor submissions across all 12 RNA targets, a model submitted as 2-5 was better than model 1 by GDT_TS, and the difference in GDT-TS between model 1 and the top scoring model for each group was no lower than if model 1 had been randomly selected (**Supplemental Figure 6A**).

Further inspection of TS110_2 helps illustrate the requirement of paying attention to the non-Watson-Crick pairs beyond the standard Watson-Crick pairs of the secondary structure, both in prediction and in assessment of RNA targets resolved by cryo-EM. Figures **7E** and **7F** show the 2D structures for R1149 as derived from the cryo-EM map (**Figure 7E**) and the best RMSD model TS110_2 (**Figure 7F**) structures. The region within the black ellipse (**Figure 7G**) contains a GU and a UU pair, but in the modeled structure, only the GU pair is reproduced and, while the right Us face each other, they do not form a pair (**Figure 7H**). In the region circled in red, the fold of the single-stranded loop is missed and in the one circled in green, the fold leads to several bad contacts between residues, which may explain the rather high clashscore of 31 for TS110_2, despite the overall good fit in the relative orientations between the helices (**Figure 7A**). It is important to note that for these regions, alternative structures in the experimentalists' 10-model cryo-EM ensemble show breaking of the features, similar to the prediction TS110_2; and so it is possible that the conformations modeled in TS110_2 occur in the actual cryo-ensemble for the target R1149. Nevertheless, these model discrepancies lead to deviations of the strands in the four-way junction that, in turn, lead to variations in the arms at the junction (**Figure 7G**). Indeed, all 10 members of the experimental cryo-EM ensemble show complete base pairing at the molecule's central four-way junction, which is inconsistent with incomplete junction base pairing in TS110_2 (**Figure 7F**).

The presence of alternative structures, noted above for the non-natural six helix bundle R1138, was a common theme in RNA targets in CASP (**Table 1**), and was particularly interesting in one target with continuous heterogeneity. R1156 is a homolog of the same SARS-CoV-2 SL5 domain as R1149, and showed flexibility in one helix (blue, **Figure 7H-I**), which was represented in the cryo-EM analysis as four sub-classified maps. Comparing models directly to these experimental maps highlighted models of particular excellent quality that fit into the maps nearly as well as the target models (**Supplemental Figure 4**). In particular, the model TS128_5 from GeneSilico, fit into the experimental map with excellent scores (**Figure 7H-I**). Fitting the model into the highest resolution of these four maps, conformation 1, we can see visually and numerically, that the model fits well with respect to 3 helices but poorly with respect to the flexible helix (**Figure 7H**). However, the model fits better in the second conformation, obtaining map-to-model atomic inclusion scores comparable to scores achieved by models derived with knowledge of the map (**Figure 7I**). This comparison revealed the importance of representing the ensemble of structures the RNA can form so as to not penalize prediction of structures that do form but can not be captured by a single experimental structure.

In summary, inspection of top-ranked CASP15 RNA models confirms, in each case, good prediction of global topology but also reveals fine details and/or aspects of conformational

heterogeneity that have not been captured by the models. Ordering each set of five models by the predictor groups also typically did not correlate with the models' accuracy. Similar conclusions for R1108, R1128, R1138, R1149 and R1156 based on alternative analyses by RNA experimental groups, are described in a separate paper being prepared for the CASP issue.

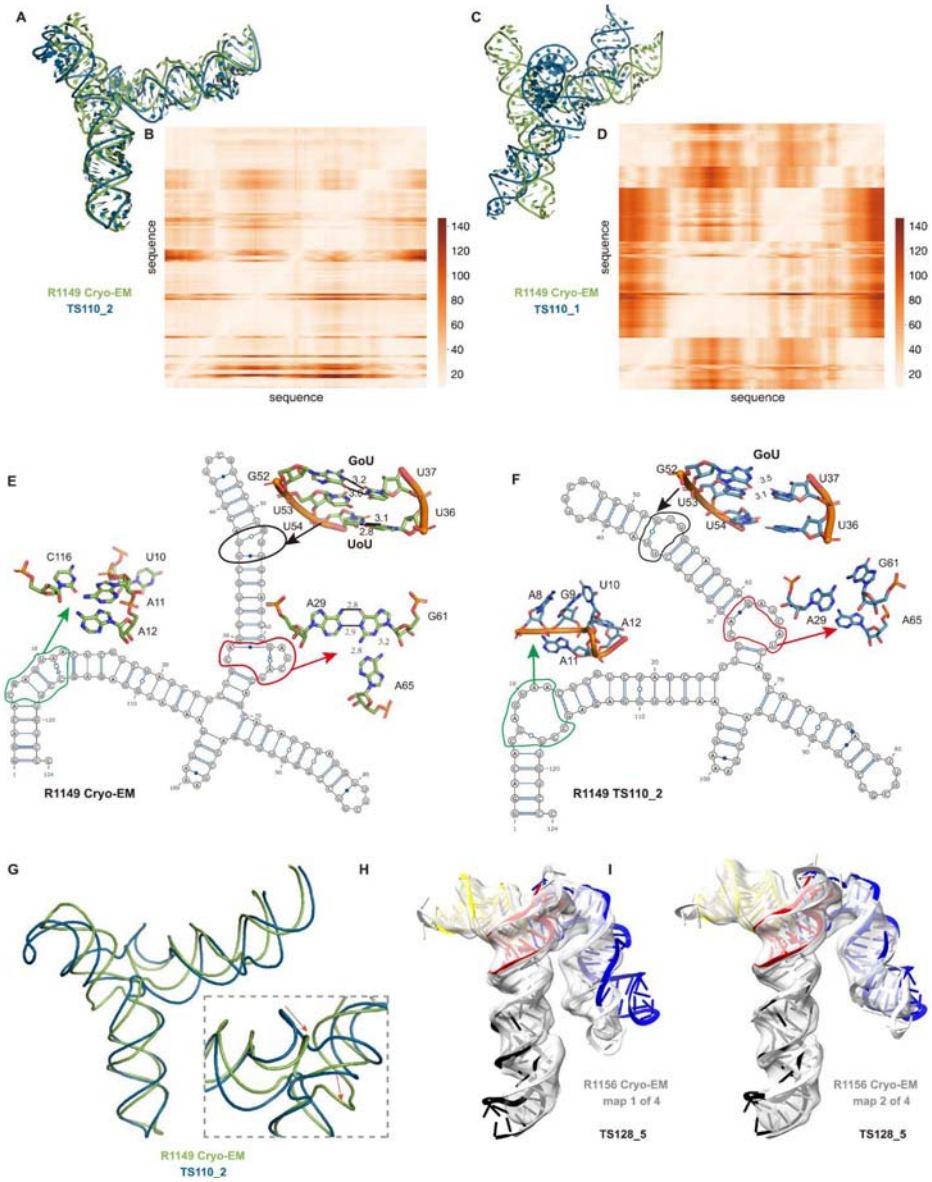

FIGURE 7. Detailed inspection of "difficult" targets, two coronavirus SL5 domains solved by cryo-EM. (A) Superposition between R1149 cryo-EM structure (first of 10 models representing experimental uncertainty) and the closest CASP15 prediction according to RMSD (TS110_2 with 6.9 Å). **(B)** Deformation profile between the same two structures. **(C)** Superposition between the experimental (R1149) and the model ranked #1 by the modeling group (TS110_1 with 21.7 Å). **(D)** Deformation profile between the same two structures. **(E)** Diagram of the secondary structure (2D) of target R1149 (first of 10 models representing experimental uncertainty). **(F)** Diagram of the secondary structure (2D) of the closest model TS110_2. The outlines indicate regions with large discrepancies due to wrong 2D pairs and absence of 3D pairs. For example, in the model structure, the U54/U36 pair is not present, and the region

circled in green shows a region with high clashscore. **(G)** Backbone traces of the experimental (green) and model (blue) structures showing the overall fit of the helices; however, as shown in inset, the wrong choices in internal loops lead to large deviations in the path of the backbone at the central 4-way junction. **(H-I)** Experimental maps and models (gray) for R1156, whose cryo-EM data was subclassified into four separate conformations; conformation 1 **(H)** and 2 **(I)** compared to top scoring CASP prediction TS128_5 (color).

3.8 Potential utility of RNA models for molecular replacement

The general topological accuracy of the CASP15 RNA tertiary structure models motivated us to explore their potential utility for phasing X-ray diffraction data by molecular replacement, which has previously been carried out in very few cases[67]. While we began these explorations in studies described above to rank models based on agreement with X-ray data (**Section 3.6**), such scores based on optimal placements do not necessarily reflect models' value as search models for real-world Molecular Replacement (MR). For example, a largely accurate model may prove unsuccessful if inaccurately modeled portions lead to severe crystal lattice packing clashes.

We therefore carried out more realistic MR runs, first, on all unmodified models of R1117. This initial analysis was restricted to R1117 since visual examination and LLG calculations suggested that models of other targets would require some kind of editing to succeed (see next). Across the up to five models submitted by groups, we found that 3 out of 34 groups succeeded with at least one model, using global map correlation coefficient CC > 0.2 as the criterion of success (**Supplemental Figure 7**). Among these successful groups, however, the quality of MR solutions varied significantly. The highest LLG was 110 for model 128_2 but results in a poor R_{free} of 54% after refinements in Refmac5; in contrast, the lowest R_{free} was 39% for model 287_3 from the Chen group after refinement with Refmac5[68] despite this model giving a worse LLG in the MR trials. **Figure 8** shows the successful solution with model 287_3.

For the other three CASP targets solved by X-ray diffraction, visual inspection and the Z_{mx} values in **Figure 5** made clear that editing of the predictions would be required for successful MR, and, to focus resources, models from TS232 (AIchemy_RNA2) were subjected to various editing procedures. For solution of the two CPEB3 ribozymes, R1107 (one protein chain, one RNA chain) and R1108 (two protein chains, two RNA chains), the structural variance observed in group TS232 models after structural alignment with Theseus [69] was used as an indication of local prediction reliability and divergent regions removed before the edited model_1 was used as a search model. This approach borrows from that taken for proteins by the MR pipeline AMPLE[70]. R1107 was successfully solved by first placing the protein chain then the edited RNA search model, both with Phaser. The result (**Figure 8**) has an R_{free} of 26% and visible density for the missing part of the RNA molecule confirms that it could be readily refined and completed. R1108, a close homolog of R1107, proved much more difficult to solve, perhaps owing to the different conformations observed between the two RNA chains in the asymmetric unit. When attempting to solve this structure similarly (protein first then RNA) we could place the protein component, but the RNA component was reversed, providing only a partial solution. The

truncated group TS232 models for R1108 were of a sufficient quality to solve R1107 and the resulting protein/RNA complex could then be used to solve R1108 with an R_{free} of 41%.

Inspection of the Group 232 models for R1116 showed that more extensive model editing would be required. A modified version of Slice'N'Dice [49] was therefore used to split model_1 into three structural units. A portion comprising nucleotides 1-24;125-157 could then be placed with Molrep which indicated a partial solution after refinement (R_{fact} 48%, R_{free} 52%). Three copies of a second fragment comprising 38-63 could then be placed to largely complete the structure with Phaser scores of LLG: 1324 and TFZ: 9.6. The result (**Figure 8**) has an R_{free} of 43%, an acceptable value for a model immediately after MR. These results demonstrate that all of the RNA crystal structure targets in CASP15 could, one way or another, be solved by MR, although it is recognised that further refinement and completion (not attempted here) could be challenging, especially at 3.0 Å or worse resolution.

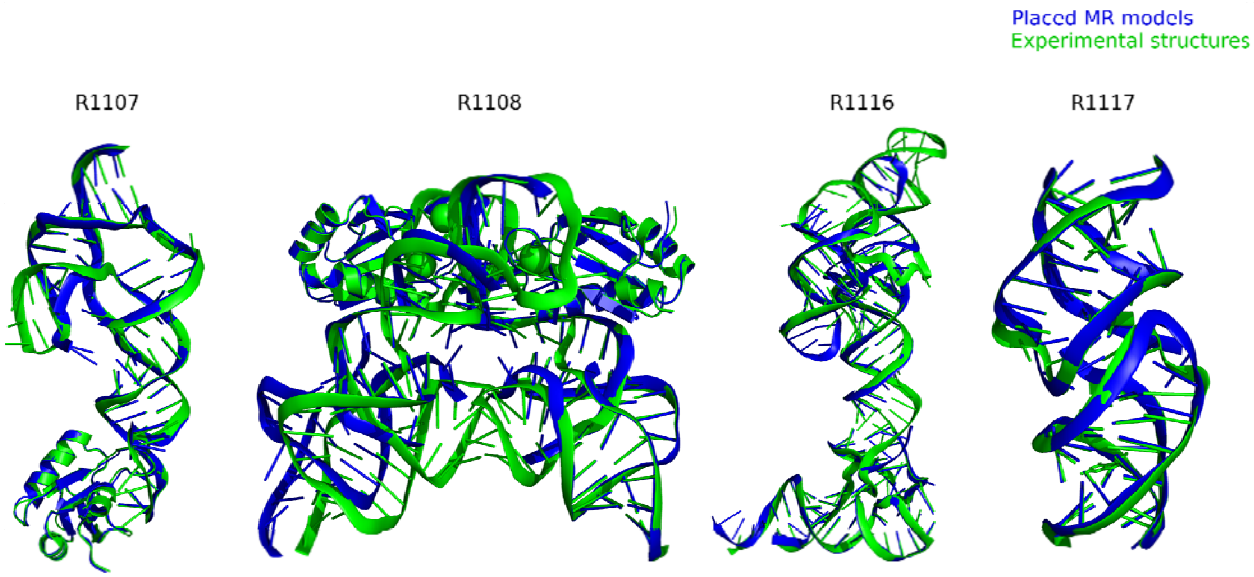

FIGURE 8. Molecular replacement (MR) of X-ray crystallographic data using CASP15 models (and AlphaFold 2 models of U1ABD in the cases of R1107 and R1108). Group TS232 models formed the basis of all successful search models shown except R1117 (group TS287).

4 Discussion

CASP15 enabled a timely assessment of 3D RNA structure prediction, with 8 RNA targets solved by cryo-EM and 4 by X-ray crystallography. Forty two predictors from 25 research centers made submissions for at least one of these targets, many of whom had not published studies on RNA prior to CASP but explored deep learning approaches that were novel for the RNA field. The twelve RNA targets ranged in difficulty from "easy", with clearly identifiable templates in the structure database, to "difficult", with no templates. When looking at all five submissions for each target, visually good predictions were submitted for all 10 RNA-only

targets, including 4 non-natural RNA targets that had no global homology to previously solved structures. Two protein-RNA complexes were not modeled accurately.

Quantitative rankings of predictor groups were carried out by independent teams, based on RMSD and INF metrics developed in the RNA-Puzzles trials and based on TM-score, GDT_TS, and lDDT more familiar to protein structure assessments in prior CASP. Both rankings agreed in placing TS232 (AIchemy_RNA2) first, TS287 (Chen) second, and TS128 and TS081 (GeneSilico and RNAPolis) as tied for third. These rankings were also confirmed by analyses comparing predicted models to maps (for cryo-EM targets) and statistics related to molecular replacement (for X-ray crystal targets). The top-ranked models for the 10 RNA-only targets captured topology well, as assessed by visual inspection and by achievement of GDT_TS values greater than 45 and/or TM-score values greater than 0.45. Nevertheless, fine details such as non canonical pairs and hydrogen bonding at junctions were inaccurate in these models, even when taking into account sources of uncertainty for the experimental structures. Conformational heterogeneity in some targets, such as presence of multiple structures captured by cryo-EM, were not captured by any group in their range of submitted models (**Supplemental Figure 8**). Despite these caveats, the general topological accuracy for RNA-only targets — even those without homologs of known structure — and the ability of models, with some curation, to enable molecular replacement of all 4 X-ray diffraction data sets suggest reason for optimism.

Has there been improvement in RNA modeling in CASP15 compared to prior RNA-puzzles? Achieving accurate positioning of helices with respect to each other by modeling is often feasible when the single-stranded segments are short and unpaired because RNA helices are bulky and the interconnecting strands each have a polarity. This good helix positioning was also regularly observed during previous RNA-Puzzles assessments and in previous RNA modeling efforts. During this CASP15 experiment, some research groups tried to use prediction approaches that were similar to AlphaFold or AlphaFold 2, both of which are AI-based methods for predicting the structure of proteins. For example, AlChemy_RNA[71] uses an end-to-end differentiable network inspired by AlphaFold 2. However, these AI-based predictions did not perform as well as expected and did not surpass prediction methods previously tested in RNA-Puzzles (SimRNA, Chen, RNApolis), which have been continuously improving for the past decade. The AI-based approaches[71,72] also failed to demonstrate the accuracy claimed in their preprint papers, perhaps due to the limited amount of training data.

In addition to not using deep learning, the top four RNA predictors shared the property that they were not servers and, based on their own accounts (see papers co-submitted for this CASP issue), they appeared to still make use of human intuition. While there were cases in lower ranked predictor groups where server models were more accurate than ‘human’ models from the same predictor group (e.g., Yang), generally server models were worse in quality than the top 4 human predictor groups. Going forward, an important frontier for the RNA structure prediction field to focus on will be automation, so that methods can be more widely used and applied at the genomic scale, as is now the case for protein structure prediction methods. While the sparser data available for RNA structure, compared to protein structure, has complicated

development of robust deep learning algorithms, recent accelerations in RNA structure determination – particularly from cryo-EM[12] – and the availability of high-throughput sequencing-based methods sensitive to RNA structure[73] may help close the gap between RNA and protein computational methods.

In addition to being the first CASP experiment for RNA structure prediction, CASP15 was also the first CASP experiment for RNA structure assessment, and future CASP RNA trials can benefit from some lessons learned by the assessors, three of which we discuss here. First, CASP15 included few truly difficult RNA targets, and these were solved by cryo-EM at resolutions worse than 3 Å. It will be important for upcoming CASP competitions to bring in experimental groups solving natural RNA targets without previously solved homologs at near-atomic resolution. Such molecules are being discovered and structurally characterized at increasing frequency, particularly for biologically interesting RNA-protein complexes.

Second, while only 2 of 12 targets in CASP15 were RNA-protein complexes, it seems feasible CASP16 will involve more RNA-protein complexes given their biological importance and amenability to cryo-EM. For assessment, it will therefore be increasingly important to develop quantitative metrics that make sense across RNA, protein, and RNA-protein interfaces. We found here that classic metrics for protein structure accuracy assessment, GDT_TS and TM-score, were useful in ranking RNA models, but their values for visually excellent RNA models seemed anomalously low for large and small targets, respectively. More local measures of accuracy, like lDDT or other assessments of contact accuracy, appeared useful here for both RNA and RNA-protein targets. These more local measures may be less affected by length variation and also more robust to fluctuations that appear common in large, extended RNA structures. The recent availability of lDDT for RNA may allow more testing of this metric in continuous trials like CAMEO and RNA-Puzzles before the next CASP.

Third, many and perhaps most of the CASP15 RNA targets showed conformational flexibility, e.g., as evidenced by differences in conformations of different monomers in crystallographic asymmetric units or, in cryo-ensembles captured by electron microscopy, by classes of conformations separable by automated subclassification and/or 3D variability analysis[36]. In the current assessment, predictor groups were scored based on the best observed agreement of all their submitted models vs. all available experimental models, effectively assuming that modelers were predicting single structures. Modeling of the full ensemble nature of these RNA systems was neither incentivized nor assessed. In future CASPs, acceptance of multi-model ensembles (with e.g., 100's or 1000's of models within each of 5 ensembles), rather than separate single-structure models, would better incentivize development of methods for predicting conformational ensembles of macromolecules, including molecular dynamics methods that have been previously difficult to assess. Furthermore, scoring of these ensembles directly against data should be feasible; e.g., log-likelihood frameworks and GPU-enabled software[74] might enable predicted multi-model cryo-ensembles to be compared to the entire collection of electron micrographs collected for a target.

Acknowledgements

We thank Ebbe Andersen for sharing in-house models of RNA designs; Nick Grishin and Lisa Kinch for advice on Z-score computations; Marta Szachniuk and Maciej Antczak for advice and numerical cross-checks for RNA assessment metrics; Gabriel Studer for extending lDDT to RNA; Marcin Magnus for advice on RNA model cleanup and assessment tools; Adam Zemla for extending LGA to RNA; Chengxin Zhang for advice on TM-score calculations; the RNA-puzzles modeling community for their participation as predictors; all experimentalists who contributed the RNA structures; and Andriy Kryshtafovych, Krzystof Fidelis, and John Moult for the invitation and dedicated support to bring RNA model assessment to CASP. This research was supported by Stanford Bio-X (to R.D. and R.C.K.); Stanford Gerald J. Lieberman Fellowship (to R.R.); the National Institutes of Health (R35 GM122579 to R.D.), the Howard Hughes Medical Institute (HHMI, to R.D.); Biotechnology and Biological Sciences Research Council (BBSRC) grant BB/S007105/1 (D.J.R.); Leibniz ScienceCampus InterACt grant W6/2018, funded by the BWFGB Hamburg and the Leibniz Association (M.T.); the Natural Science Foundation of China (32270707 to Z.M.), the National Key R&D Program of China (2021YFF1200900, 2021YFF1200903 to Z.M.); R&D Program of Guangzhou Laboratory (Grant No. SRPG22-003, SRPG22-006, SRPG22-007 to Z.M.); French National Research Agency (LABEX: ANR-10-LABX-0036_NETRNA, Investments for the Future program, to E.W.). This article is subject to HHMI's Open Access to Publications policy. HHMI lab heads have previously granted a nonexclusive CC BY 4.0 license to the public and a sublicensable license to HHMI in their research articles. Pursuant to those licenses, the author-accepted manuscript of this article can be made freely available under a CC BY 4.0 license immediately upon publication.

Author contributions

R.D., R.C.K., and E.W. coordinated the research. Z.M. and E.W. and, independently, R.D., R.C.K., P.P., and R.R. designed and conducted global model-to-model RNA assessments. T.M. and M.T. designed and conducted protein-RNA topology assessment. R.C.K. designed and conducted cryo-EM map-to-model assessment. A.S., R.K., and D.R. designed and conducted molecular replacement tests. R.D., R.C.K., A.S., T.M., P.P., R.R., and E.W. drafted the paper and prepared figures. F.B. prepared figures; and Z.M., P.P., F.B., and R.C.K. designed and curated code repositories.

Conflict of Interest

All authors declare that they have no competing interests.

References

[1] Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, et al. STRUCTURE OF A RIBONUCLEIC ACID. *Science*. 1965;147(3664):1462-1465.

[2] Madison JT, Everett GA, Kung H. Nucleotide sequence of a yeast tyrosine transfer RNA. *Science*. 1966;153(3735):531-534.

[3] Fuller W, Hodgson A. Conformation of the anticodon loop intRNA. *Nature*. 1967;215(5103):817-821.

[4] Levitt M. Detailed molecular model for transfer ribonucleic acid. *Nature*. 1969;224(5221):759-763.

[5] Hingerty B, Brown RS, Jack A. Further refinement of the structure of yeast tRNAPhe. *J Mol Biol*. 1978;124(3):523-534.

[6] Sussman JL, Holbrook SR, Warrant RW, Church GM, Kim SH. Crystal structure of yeast phenylalanine transfer RNA. I. Crystallographic refinement. *J Mol Biol*. 1978;123(4):607-630.

[7] Westhof E, Leontis NB. An RNA-centric historical narrative around the Protein Data Bank. *J Biol Chem*. 2021;296:100555.

[8] Cruz JA, Blanchet MF, Boniecki M, Bujnicki JM, Chen SJ, Cao S, et al. RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA*. 2012;18(4):610-625.

[9] Miao Z, Adamiak RW, Antczak M, Batey RT, Becka AJ, Biesiada M, et al. RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA*. 2017;23(5):655-672.

[10] Miao Z, Adamiak RW, Antczak M, Boniecki MJ, Bujnicki J, Chen SJ, et al. RNA-Puzzles Round IV: 3D structure predictions of four ribozymes and two aptamers. *RNA*. 2020;26(8):982-995.

[11] Miao Z, Adamiak RW, Blanchet MF, Boniecki M, Bujnicki JM, Chen SJ, et al. RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA*. 2015;21(6):1066-1084.

[12] Das R. RNA structure: a renaissance begins? *Nat Methods*. 2021;18(5):439.

[13] 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.

[14] Hogan MJ, Pardi N. mRNA Vaccines in the COVID-19 Pandemic and Beyond. *Annu Rev Med*. 2022;73:17-39.

[15] Lensink MF, Brysbaert G, Mauri T, Nadzirin N, Velankar S, Chaleil RAG, et al. Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins*. 2021;89(12):1800-1823.

[16] Lensink MF, Brysbaert G, Nadzirin N, Velankar S, Chaleil RAG, Gerguri T, et al. Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. *Proteins*. 2019;87(12):1200-1221.

[17] Lensink MF, Velankar S, Baek M, Heo L, Seok C, Wodak SJ. The challenge of modeling protein assemblies: the CASP12-CAPRI experiment. *Proteins*. 2018;86 Suppl 1:257-273.

[18] Lensink MF, Velankar S, Kryshtafovych A, Huang SY, Schneidman-Duhovny D, Sali A, et al. Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins*. 2016;84 Suppl 1(Suppl Suppl 1):323-348.

[19] Parisien M, Cruz JA, Westhof E, Major F. New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*. 2009;15(10):1875-1885.

[20] Gendron P, Lemieux S, Major F. Quantitative analysis of nucleic acid three-dimensional structures. *J Mol Biol*. 2001;308(5):919-936.

[21] Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*. 1975;405(2):442-451.

[22] Gorodkin J, Stricklin SL, Stormo GD. Discovering common stem–loop motifs in unaligned RNA sequences. *Nucleic Acids Res*. 2001;29(10):2135-2144.

[23] Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res*. 2007;35(Web Server issue):W375-W383.

[24] Murray LJW, Arendall WB 3rd, Richardson DC, Richardson JS. RNA backbone is rotameric. *Proc Natl Acad Sci U S A*. 2003;100(24):13904-13909.

[25] Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, et al. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J Mol Biol*. 1999;285(4):1711-1733.

[26] Gong S, Zhang C, Zhang Y. RNA-align: quick and accurate alignment of RNA 3D structures based on size-independent TM-scoreRNA. *Bioinformatics*. 2019;35(21):4459-4461.

[27] Zok T, Popenda M, Szachniuk M. MCQ4Structures to compute similarity of molecule structures. *CEJOR Cent Eur J Oper Res*. 2014;22(3):457-473.

[28] Zhang C, Shine M, Pyle AM, Zhang Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat Methods*. 2022;19(9):1109-1115.

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

[30] Magnus M, Antczak M, Zok T, Wiedemann J, Lukasiak P, Cao Y, et al. RNA-Puzzles toolkit: a computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools. *Nucleic Acids Res*. 2020;48(2):576-588.

[31] Waleń T, Chojnowski G, Gierski P, Bujnicki JM. ClaRNA: a classifier of contacts in RNA 3D structures based on a comparative analysis of various classification schemes. *Nucleic Acids Res*. 2014;42(19):e151.

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

[33] Biasini M, Mariani V, Haas J, Scheuber S, Schenk AD, Schwede T, et al. OpenStructure: a flexible software framework for computational structural biology. *Bioinformatics*. 2010;26(20):2626-2628.

[34] Liebschner D, Afonine PV, Baker ML, Bunkóczi G, Chen VB, Croll TI, et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in

Phenix. *Acta Crystallogr D Struct Biol*. 2019;75(Pt 10):861-877.

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

[36] Beton JG, Cragnolini T, Kaleel M, Mulvaney T, Sweeney A, Topf M. Integrating model simulation tools and cryo electron microscopy. *Wiley Interdiscip Rev Comput Mol Sci*. Published online November 21, 2022. doi:10.1002/wcms.1642

[37] Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci*. 2021;30(1):70-82.

[38] Liebschner D, Afonine PV, Baker ML, Bunkóczi G, Chen VB, Croll TI, et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol*. 2019;75(Pt 10):861-877.

[39] Cragnolini T, Sahota H, Joseph AP, Sweeney A, Malhotra S, Vasishtan D, et al. TEMPy2: a Python library with improved 3D electron microscopy density-fitting and validation workflows. *Acta Crystallogr D Struct Biol*. 2021;77(Pt 1):41-47.

[40] Lagerstedt I, Moore WJ, Patwardhan A, Sanz-García E, Best C, Swedlow JR, et al. Web-based visualisation and analysis of 3D electron-microscopy data from EMDB and PDB. *J Struct Biol*. 2013;184(2):173-181.

[41] Pintilie G, Zhang K, Su Z, Li S, Schmid MF, Chiu W. Measurement of atom resolvability in cryo-EM maps with Q-scores. *Nat Methods*. 2020;17(3):328-334.

[42] McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. Phaser crystallographic software. *J Appl Crystallogr*. 2007;40(Pt 4):658-674.

[43] Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr*. 2011;67(Pt 4):235-242.

[44] Krissinel E, Lebedev AA, Uski V, Ballard CB, Keegan RM, Kovalevskiy O, et al. CCP4 Cloud for structure determination and project management in macromolecular crystallography. *Acta Crystallogr D Struct Biol*. 2022;78(Pt 9):1079-1089.

[45] Vagin A, Teplyakov A. Molecular replacement with MOLREP. *Acta Crystallogr D Biol Crystallogr*. 2010;66(Pt 1):22-25.

[46] Theobald DL, Wuttke DS. THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics*. 2006;22(17):2171-2172.

[47] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589.

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

[49] Simpkin AJ, Elliott LG, Stevenson K, Krissinel E, Rigden D, Keegan RM. Slice'N'Dice: Maximising the value of predicted models for structural biologists. *bioRxiv*. Published online

July 2, 2022. doi:10.1101/2022.06.30.497974

[50] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Published online 2012. doi:10.48550/ARXIV.1201.0490

[51] Salehi-Ashtiani K, Lupták A, Litovchick A, Szostak JW. A genomewide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene. *Science*. 2006;313(5794):1788-1792.

[52] Schroeder GM, Cavender CE, Blau ME, Jenkins JL, Mathews DH, Wedekind JE. A small RNA that cooperatively senses two stacked metabolites in one pocket for gene control. *Nat Commun*. 2022;13(1):199.

[53] Jenkins JL, Krucinska J, McCarty RM, Bandarian V, Wedekind JE. Comparison of a preQ1 riboswitch aptamer in metabolite-bound and free states with implications for gene regulation. *J Biol Chem*. 2011;286(28):24626-24637.

[54] Read RJ, Chavali G. Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins*. 2007;69 Suppl 8(S8):27-37.

[55] 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.

[56] Haas J, Barbato A, Behringer D, Studer G, Roth S, Bertoni M, et al. Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins*. 2018;86:387-398.

[57] Robin X, Haas J, Gumienny R, Smolinski A, Tauriello G, Schwede T. Continuous Automated Model EvaluatiOn (CAMEO)-Perspectives on the future of fully automated evaluation of structure prediction methods. *Proteins*. 2021;89(12):1977-1986.

[58] Przytula-Mally AI, Engilberge S, Johannsen S, Olieric V, Masquida B, Sigel RKO. Anticodon-like loop-mediated dimerization in the crystal structures of HdV-like CPEB3 ribozymes. *bioRxiv*. Published online September 22, 2022:2022.09.22.508989. doi:10.1101/2022.09.22.508989

[59] Vallina NS, McRae EKS, Hansen BK, Boussebayle A, Andersen ES. RNA origami scaffolds as a cryo-EM tool for investigating aptamer-ligand binding of a Broccoli-Pepper FRET pair. *bioRxiv*. Published online August 25, 2022:2022.08.25.505116. doi:10.1101/2022.08.25.505116

[60] McRae EKS, Rasmussen HØ, Liu J, Bøggild A, Nguyen MTA, Sampedro Vallina N, et al. Structure, folding and flexibility of co-transcriptional RNA origami. *Nat Nanotechnol*. Published online 2023:1-10.

[61] Cragnolini T, Kryshtafovych A, Topf M. Cryo-EM targets in CASP14. *Proteins*. 2021;89(12):1949-1958.

[62] Vasishtan D, Topf M. Scoring functions for cryoEM density fitting. *J Struct Biol*. 2011;174(2):333-343.

[63] Joseph AP, Lagerstedt I, Patwardhan A, Topf M, Winn M. Improved metrics for comparing

structures of macromolecular assemblies determined by 3D electron-microscopy. *J Struct Biol*. 2017;199(1):12-26.

[64] Leontis NB, Westhof E. Geometric nomenclature and classification of RNA base pairs. *RNA*. 2001;7(4):499-512.

[65] Bonilla SL, Vicens Q, Kieft JS. Cryo-EM reveals an entangled kinetic trap in the folding of a catalytic RNA. *Sci Adv*. 2022;8(34):eabq4144.

[66] Li S, Palo MZ, Pintilie G, Zhang X, Su Z, Kappel K, et al. Topological crossing in the misfolded Tetrahymena ribozyme resolved by cryo-EM. *Proc Natl Acad Sci U S A*. 2022;119(37):e2209146119.

[67] Huang L, Wang J, Watkins AM, Das R, Lilley DMJ. Structure and ligand binding of the glutamine-II riboswitch. *Nucleic Acids Res*. 2019;47(14):7666-7675.

[68] Murshudov GN, Skubák P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, et al. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr*. 2011;67(4):355-367.

[69] Theobald DL, Wuttke DS. Accurate Structural Correlations from Maximum Likelihood Superpositions. *PLoS Comput Biol*. 2008;4(2):e43.

[70] Bibby J, Keegan RM, Mayans O, Winn MD, Rigden DJ. AMPLE: a cluster-and-truncate approach to solve the crystal structures of small proteins using rapidly computed ab initio models. *Acta Crystallogr D Biol Crystallogr*. 2012;68(Pt 12):1622-1631.

[71] Shen T, Hu Z, Peng Z, Chen J, Xiong P, Hong L, et al. E2Efold-3D: End-to-End Deep Learning Method for accurate de novo RNA 3D Structure Prediction. Published online July 4, 2022. http://arxiv.org/abs/2207.01586. Accessed April 22, 2023

[72] Pearce R, Omenn GS, Zhang Y. De Novo RNA Tertiary Structure Prediction at Atomic Resolution Using Geometric Potentials from Deep Learning. *bioRxiv*. Published online May 15, 2022:2022.05.15.491755. doi:10.1101/2022.05.15.491755

[73] Strobel EJ, Yu AM, Lucks JB. High-throughput determination of RNA structures. *Nat Rev Genet*. 2018;19(10):615-634.

[74] Cossio P, Rohr D, Baruffa F, Rampp M, Lindenstruth V, Hummer G. BioEM: GPU-accelerated computing of Bayesian inference of electron microscopy images. *Comput Phys Commun*. 2017;210:163-171.