Limits of deep-learning-based RNA prediction methods

Marko Ludaic, Arne Elofsson*

Science for Life Laboratory and Department of Biochemistry and Biophysics, Stockholm University, Solna 171 21, Sweden
*Corresponding author. Department of Biochemistry and Biophysics and Science for Life Laboratory, Stockholm University, Box 1031, 171 21 Solna, Sweden.

Abstract

***Motivation**:* In recent years, tremendous advances have been made in predicting protein structures and protein-protein interactions. However, progress in predicting the structure of RNA, either alone or in complex with other macromolecules, has been less prominent, although some recent developments have been reported. It remains unclear whether the improved prediction accuracy is sustained for novel RNA structures.

Results: Here, we use an independent benchmark to evaluate the performance of the latest methods. First, we show that state-of-the-art methods can sometimes predict the structure of single-chain RNA strands, with accurate models observed for RNAs with well-defined or regular secondary structures. Next, our evaluation was extended to RNA complexes, where prediction accuracy was notably higher for those involving extensive canonical base pairing. Additionally, a structural similarity analysis revealed that prediction success strongly correlates with resemblance to known structures, indicating that current methods recognise recurring motifs rather than generalising to novel folds. Finally, we also noted that the estimation of accuracies for RNA models is far from accurate. Therefore, it is not possible to reliably identify the correctly predicted models with today's methods.

***Availability and Implementation**:* Code used for running the methods and analyzing the main results is available from https://github.com/iammarcol/RNA-Benchmark and here https://zenodo.org/records/15304777

***Contact**: arne@bioinfo.se*

***Supplementary information**: Supplementary Figure S1-S14, Supplementary Tables S1-S5*

Keywords: RNA, RNA-protein interactions, AlphaFold3, benchmark, machine learning

1 Introduction

RNA molecules are highly dynamic macromolecules that adopt intricate three-dimensional (3D) structures essential for many cellular and viral processes, emergent RNA-based medicines, and biotechnologies. Different types of RNA, such as antisense oligonucleotides, small interfering RNAs, and microRNAs, have shown promising results as therapeutics (Damase *et al.*, 2021). Even though RNAs were initially viewed as a mere intermediary in transferring genetic information from DNA to protein synthesis, they have been recently recognised for their functional versatility, extending far beyond their linear nucleotide sequence.

RNA structures are critical for elucidating structure-function relationships; therefore, modelling RNA structure from its nucleotide sequences remains an important and largely unsolved task. There are two main reasons for the limited success. First, the limited availability of determined RNA structures, at least one order of magnitude less than for proteins, poses a significant challenge for training AI-based models. Secondly, RNA's conformational dynamicity complicates the application of deep learning models. Many RNAs undergo drastic conformational changes to obtain a shape required for binding to a molecular partner (Ma *et al.*, 2022). This is also one of the reasons why experimental methods such as X-ray crystallography, NMR, and cryo-EM have been primarily able to resolve regulatory and enzymatic RNAs such as tRNA, rRNA, and ribozymes (Zhang *et al.*, 2022). The structure of other RNA types like long non-coding RNAs (lncRNAs) remains less studied, possibly because they are indeed more disordered.

Prior to the release of RoseTTAFold2NA (RF2NA) in early 2024 (Baek *et al.*, 2024) and AlphaFold3 later in 2024 (Abramson *et al.*, 2024), deep learning approaches were limited to predicting the structures of monomeric RNAs and could not handle RNA in a complex with other molecules. Recently, several deep learning approaches based on the AlphaFold3 architecture, such as Boltz-1 (Wohlwend *et al.*, 2024), Chai-1 (Discovery *et al.*, 2024), and HelixFold3 (Liu *et al.*, 2024), were released. These methods can handle proteins, nucleic acids, small molecules, and ions, greatly expanding the scope and applicability of structure prediction. Furthermore, some methods for predicting single-chain RNA, such as RhoFold+ (Shen et al., 2024), utilize large language models (LLMs) to predict atom coordinates, while others are adaptations of AlphaFold. Table 1 provides an overview of the latest methods for predicting RNA structure.

Since advancements in RNA structure prediction are recent, only one benchmark study focused on a single-chain RNA has been reported previously (Bernard *et al.*, 2024). Some benchmarking results were presented at the CASP16 meeting in December 2024, but those did not include the latest methods as they were released too late to participate in CASP16.

In this study, we conducted a comprehensive benchmark study to evaluate the capabilities and limitations of these deep learning methods for both single-chain RNA and RNA complexes. The analysis compared the performance of nine methods for predicting monomeric RNA structures and four methods for predicting RNA complexes. By systematically assessing these models, we aim to provide valuable insights into their accuracy, strengths, and areas for further improvement.

2 Methods

2.1 Benchmark dataset

Nucleotide-containing entries released after September 30, 2021, were downloaded from the biological assemblies in the Protein Data Bank (PDB) along with their corresponding sequences. The first biological unit was used. In addition, we used CASP16 (Critical Assessment of Structure Prediction) targets containing RNA to supplement the benchmark dataset. Next, sequences were filtered using CD-HIT-EST (Li and Godzik, 2006) at 80% sequence identity to ensure dataset diversity. The identity was calculated using the longer sequence as a reference, which was critical, as many RNA sequences in the PDB are synthesized from short RNAs. Applying this criterion avoids falsely reporting a sequence identity of 100% between a short and a long RNA sequence. Sequences containing ambiguous nucleotides, such as 'X' or 'N', were removed from the dataset. Furthermore, sequences shorter than 10 nucleotides that consisted predominantly of a single nucleotide type (>80% homogeneity) were also excluded, as they are likely to be linear and lack a modular structure, making them unsuitable for inclusion in the benchmark dataset.

MMseqs2 (Steinegger and Söding, 2017) was employed to assess sequence homology between proteins in the benchmark dataset and those in the AlphaFold training dataset (PDB structures released before September 30, 2021). For each RNA-protein complex, if at least one protein chain shared more than 40% sequence identity with any sequence in the AlphaFold training dataset, the corresponding RNA molecule in the same complex was also analyzed for sequence identity. Only if it had more than 80% sequence identity, the complex was excluded, since the primary focus of this benchmark was to evaluate the accuracy of RNA prediction methods, rather than protein prediction methods.

Finally, the dataset was reduced as we could not model some entries in the initial test set for some methods. Modelling complexes with more than five protein and/or RNA subunits with Boltz-1 version 0.4.1 failed due to out-of-memory errors. Consequently, such complexes were excluded from the analysis, reducing the RNA-RNA and RNA-protein complexes dataset to 42 examples. In the case of single-chain RNA structures, after filtering, our benchmark dataset consisted of 50 commonly generated single-chain RNA models.

2.2 Model generation

The benchmark evaluates the performance of nine prediction methods, AlphaFold3 (Abramson *et al.*, 2024), Boltz-1 (Wohlwend *et al.*, 2024), Chai-1 (Discovery *et al.*, 2024), HelixFold3 (Liu

et al., 2024), NuFold (Kagaya *et al.*, 2025), RhoFold+ (Shen *et al.*, 2024), RoseTTAFoldNA (Baek *et al.*, 2024), and trRosettaRNA (Wang *et al.*, 2023), for single-chain RNA. Four of the methods. AlphaFold3, Boltz-1, HelixFold3, and RoseTTAFoldNA also predicted RNA-RNA and RNA-protein complexes. Chai-1, a multi-modal foundation model for molecular structure prediction, was excluded from the RNA complex analysis due to its inference limit of 2048 tokens.

For each method, the structures were generated using a single NVIDIA DGX-A100 Core GPU, with the latest version of each method (v1.0 for trRosettaRNA) as of December 1, 2024, and default settings. For all methods, the same set of MSAs was used. These were generated using rMSA (Zhang *et al.*, 2023) for RNA sequences and MMseqs2 for protein sequences. An rMSA search was performed against the RNACentral and NT databases using default settings. For proteins, MMseqs2 was run on the ColabFold databases as of December 1, 2024, which consist of UniRef30, BFD/Mgnify, and ColabFold DB (Mirdita *et al.*, 2022). MSAs were converted into alternative formats depending on each model's specific requirements. The conversion scripts are available from the accompanying git repository. Twenty-five models, in the case of AlphaFold3, and five models for the other methods, were generated using different seeds.

The top-ranked model was used for analyzing each method. Several methods, including AlphaFold3, report the predicted template modeling (pTM) and the interface predicted template modeling (ipTM) scores, and rank models by combining pTM and ipTM scores calculated as follows:

```
0.8*ipTM + 0.2*pTM + 0.5*disorder - 100*clash.
```

Other methods, such as RF2NA or RhoFold+, report the predicted local distance difference test (pLDDT), which reflects confidence in the local structure (Mariani et al., 2013). We ranked the models based on the average pLDDT scores for these methods. If a method did not report any confidence or ranking score, we randomly selected one of the five models. We also noted that in almost all cases, the difference in quality between generated models is minimal (see Figures S1-S4).

2.3 Model evaluation

To evaluate the quality of predicted models, we used TM-score (Zhang and Skolnick, 2005) for assessing single-chain RNA structures and the DockQ score (Basu and Wallner, 2016; Mirabello and Wallner, 2024) in addition to the TM-score for evaluating RNA complexes.

When calculating TM-score using USalign's default parameters (C. Zhang et al., 2022), we observed unexpectedly low values, even when the structures appeared to be correct. For example, a double helical RNA (8SXL) modeled with AlphaFold3, which is almost perfectly predicted (RMSD = 0.47), had a TM-score of only 0.660 (Figure S5). The discrepancy was traced to the length normalization method used by TM-score. TM-score employs a scale factor (d_0) that is dependent on the shortest chain length, L_{min}, to calculate the similarity between two residues *i* and *j,* separated by a distance d_{ij}:

$$s(i,j) = \frac{1}{1+(\frac{d_{ij}^2}{d_0})} \text{ and } d_0 = 1.24\sqrt{L_{min} - 15} - 1.8$$

This formula is optimized for proteins, but we realized that d_0 was often too low for RNA chains, as they are usually shorter than proteins. To our knowledge, no one has tried to optimize TMscore for nucleotides. We observed that using a constant d_o of 5.0 produced realistic results and did not try to optimize that further. Therefore, we calculate the TM-score using the following command line options:

```
{model.pdb} {reference.pdb} -mm 1 -d 5 -ter 0
```

for RNA complexes and

```
{model.pdb} {reference.pdb} -mol RNA -d 5
```

for single-chain RNA molecules.

Similarly, we calculate the TM-score for protein chains using the following command line option:

```
{model.pdb} {reference.pdb} -mol prot
```

Changing the d_0 parameter to 5.0 provided a more realistic TM-score for the AlphaFold3 model of 8SXL of 0.965 (Figure S5).

DockQ provides an overall score, averaged across all interfaces, as well as individual scores, one for each interface. Therefore, we also assessed the accuracy of RNA-RNA and RNA-protein interfaces within each complex, using the per-interface DockQ score.

2.4 Structural similarity to the training set

We also conducted a structural similarity analysis between the benchmark dataset and the AlphaFold3 training dataset, which is identical to the training set of most methods, to identify structural redundancy between the training and benchmark sets. Specifically, we downloaded all

RNA-containing entries, single chains and complexes, present in PDB before September 30, 2021, which corresponds to the cutoff date for the AlphaFold3 training set.

A structural search was performed for all entries in the test set against the downloaded PDB set using USalign with the same settings described above. The TM-scores were computed between each RNA structure in the training set and all the RNA-containing structures in the training set. For each entry in the benchmark dataset, the maximum TM-score in the training set represents the highest structural similarity to any structure in the training set. This number was then used to classify targets based on their structural similarity to the training set.

3 Results

3.1 Performance of structure prediction methods on single-chain RNAs

Figure 1 shows the performance of deep learning methods for single-chain RNA prediction. In addition to the benchmark set of monomeric RNAs, we evaluated the prediction accuracy of individual chains in the RNA-RNA and RNA-protein complexes for the four methods that predict complexes. Figure 1A shows the distribution of TM-scores for each method, and the success rates, i.e. the fraction of models with a TM-score ≥ 0.5, are shown in Figure 1B. TM-scores for all models are shown in Table S1. Overall, Boltz-1 demonstrates the best performance, achieving an average TM-score of 0.525 and a success rate of 56%, followed by AlphaFold3 (0.491/44%) and RF2NA (0.478/48%). A paired T-test reported no significant difference between the predictions from Boltz-1 and AlphaFold3, but NuFold (P = 8.2*10-3), HF3 (P = 1.210-3), and trRosettaRNA (P = 2.7*10-3) are statistically significantly less accurate than Boltz-1. A similar trend can be seen for the chains, predicted as a part of a complex, with the highest mean TM-score of 0.669 for RNA chains obtained by Boltz-1. As expected, the protein components are predicted better, 0.887/91% for AlphaFold3. However, no significant statistical difference was found between AlphaFold3 and Boltz-1 methods, see Table S2.

To better understand performance differences across different types of RNAs, we plotted the TM scores of the top-performing method, Boltz-1 (Figure 1C and 1D). We categorized the RNAs into short RNA loops, G-quadruplexes, L-shaped structure conformations such as those seen in tRNA, and Complex RNA structures, which are typically longer and composed of multiple distinct secondary structure motifs. First, the prediction quality is highly correlated across different methods, as shown in Figures 1C, 1D, and S6. That is, some targets are generally easier to predict than others. In general, short RNA loops, G-quadruplexes, and L-shaped structure conformations (Figure 2) are easier to predict, compared with longer and more complex RNA

structures (Figure S7). Exception to this trend is a couple of models, including two G-quadruplex structures, 7SXP and 8EDP, where Boltz-1 makes a significantly better model than AlphaFold3 and the other methods.

3.2 Performance of structure prediction for RNA-RNA and RNA-protein complexes

The performance of the benchmarked methods, evaluated using TM-score and DockQ score, is shown as violin plots in Figures 3A and 3B, respectively. Among the four methods, AlphaFold3 demonstrated the highest mean TM-score (0.635; Figure 3A) and DockQ score of (0.394; Figure 3B), followed by Boltz-1 (0.597/0.339), HF3 (0.502/0.309), and RF2NA (0.462/0.122). Using a DockQ score of ≥ 0.23 to define successful models, the success rates for benchmarked methods are 54.8%, 50.0%, 38.1%, and 19.0% for AlphaFold3, Boltz-1, HF3, and RF2NA, respectively. TM-scores and DockQ scores across all methods are shown in Tables S3 and S4, respectively. Mean TM-scores and success rates are shown in Table S5.

To gain more insights, we analyzed individual RNA-RNA and RNA-protein interfaces within each complex by calculating DockQ scores for each pair of chains, i.e., the interface DockQ. The individual DockQ scores for RNA-RNA interfaces, compared to those for RNA-protein interfaces, suggest that there is no significant difference in prediction accuracy between these two types (Tables S5 and S6). Here, AlphaFold3 shows the highest mean DockQ scores for both RNA-protein (0.382) and RNA-RNA interfaces (0.355), followed by Boltz-1 (0.309/0.317).

The relationships of the model qualities of the different methods are shown in Figure 3E, 3F, and S8. AlphaFold3 outperformed Boltz-1 in some cases, but no statistically significant difference was found between AlphaFold3 and Boltz-1 for TM-Score ($P = 7.1*10^{-2}$; Figure 3D) or interface DockQ score ($P = 1.65*10^{-1}$; Figure 3E). Mean TM-scores and DockQ scores across all methods revealed that the same model consistently achieved the highest scores. Notably, this highly accurate model is a double helical RNA homodimer (8SXL), likely benefiting from the relative simplicity of base pair complementarity, which makes it easier to predict. The structure comparison across all methods for this model, as well as another high-scoring model, 8TQX, is shown in Figure S5.

3.3 Models with low DockQ scores and high TM-scores

For most models, the DockQ score and TM-score agree well (Figure S9). However, for five RNA-protein interfaces (7VSJ, 7YEY, 8E2A, 8KI8, M1282) and one RNA-RNA interface (7XKM), we observed a high TM score (>0.5), despite a very low interface DockQ score (<0.23). In these cases, the RNA molecule was misplaced, bound to the wrong region of the larger protein chain. In two of these cases (7VSJ, 7YEY), the single protein chain has an internal

structural symmetry. The RNA chain was incorrectly placed on the opposite side of the symmetric fold, resulting in an incorrect prediction of the interface. Although the two sides are structurally similar, the sequences are not (Figure S10).

4 Discussion

As shown above, the best methods, Boltz-1 and AlphaFold3, make acceptable predictions for about half of the individual RNA chains and complexes. Here, we will first discuss how well the methods identify correct predictions and then how their performance depends on similarity to the training set.

4.1 The reliability of pTM/ipTM confidence scores

Several methods, including AlphaFold3, Boltz-1, and HelixFold3, provide confidence metrics alongside predicted structures in the form of predicted Template Modelling (pTM) scores for both monomeric and complex structures, and the interface predicted Template Modelling (ipTM) scores for macromolecular complexes. To assess the predictive reliability of these metrics, we compared them with the calculated TM-score for single-chain RNA and RNA complexes, as well as the DockQ score for RNA complexes.

For single-chain RNAs, the reported pTM scores were low (pTM < 0.5), indicating limited confidence in the structural predictions across all methods (see Figure 4). Even more striking is the low correlation between pTM and TM scores. Often, the pTM scores are below 0.5, even for correctly predicted models, and the correlation is best for HelixFold3, at 0.59. In contrast, for some complexes, particularly protein-RNA complexes, the ipTM scores are relatively high (greater than 0.8), as shown in Figure 4. The correlation is also higher, with the best correlation between ipTM and DockQ score found for HF3 (0.77), followed by Boltz-1 (0.70) and AlphaFold3 (0.58). Notably, RNA-RNA complexes have lower ipTM and DockQ scores than RNA-protein complexes, indicating that these are harder to predict. We hypothesize that the increased confidence for RNA-protein complexes is due to the presence of protein chains in the complexes.

4.3 Prediction accuracy dependence on AlphaFold3 training set

The presence of commonly found RNA motifs in the PDB raises the possibility that the benchmarked methods may perform well by recognizing frequently occurring patterns, rather than generalizing to novel or less common RNA structures. In other words, the methods may primarily be able to detect structural similarity to the training set. As our results suggest, double helical and cloverleaf structures are usually predicted with higher accuracy, but these are also

more common in PDB. To investigate the extent to which model accuracy is influenced by prior exposure to similar motifs during training, we assessed whether motifs of benchmarked structures were present in AlphaFold3's training data.

For both single RNA (Figure 5A, 5B, and S11-S13) and RNA complexes (Figure 5C, 5D, and S14), we defined four similarity categories using the following TM-score thresholds: <0.25, 0.25–0.5, 0.5–0.75, and >0.75, We observed that prediction accuracy, as measured by TM-score, generally increased with higher structural similarity to the training set for all methods. The mean TM-score is below 0.1 for the category with the least similarity and increases gradually to over 0.6 for the category with the highest similarity to the training set. This suggests that AlphaFold3 and other methods tend to perform better when the target structure is more similar to motifs it encountered during training, highlighting the limitation of current methods in predicting unseen and structurally divergent RNAs.

5 Conclusion

Advancements in deep learning have significantly impacted RNA structural biology, enabling the prediction of both monomeric RNA structures and RNA complexes. In this benchmark study, we evaluated several state-of-the-art deep learning methods for RNA structure prediction, revealing their strengths and limitations. Our results indicate that approximately one-third of the models achieve high accuracy. The accurately predicted models correspond primarily to RNA molecules with structural similarity to already known RNA structures. Given that current methods are trained on existing experimentally solved structures, our findings suggest that the lack of RNA structural diversity in the PDB hinders generalization to novel and structurally complex RNAs, and that current state-of-the-art methods are unable to generalize outside their training regime. Further, even within the training regime, the ability to reliably identify the correct models is a severe limitation for large-scale predictions.

Anyhow, RNA is increasingly recognized as a versatile macromolecule with essential catalytic and regulatory functions, many of which are unknown. We anticipate that growing interest in RNA biology will stimulate further experimental efforts towards solving complex RNA structures. Combined with innovations in techniques such as cryo-EM (Ma *et al.*, 2022) and continued advances in machine learning (Kwon, 2025), we are optimistic that RNA 3D structure prediction will evolve beyond simple motifs and toward accurately modelling the full complexity of RNA structural space. A possible first step would be to improve the estimates of model accuracy.

Acknowledgements

We thank the members of the Elofsson group for valuable input.

Author contributions

ML generated all the data and wrote the initial draft of the manuscript. AE and ML jointly analyzed and finalized the manuscript.

Supplementary data

Conflict of interest

The authors declare no conflict of interest.

Funding

AE was funded by the Vetenskapsrådet Grant No. 2021-03979 and the Knut and Alice Wallenberg Foundation, Grant No. 2022.0032. The computations and data handling were enabled by the supercomputing resource Berzelius, provided by the National Supercomputer Centre at Linköping University, the Knut and Alice Wallenberg Foundation, and SNIC, with grant Numbers. SNIC 2021/5-297 and Berzelius-2021-29.

References

Abramson,J. *et al.* (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, **630**, 493–500.

Baek,M. *et al.* (2024) Accurate prediction of protein-nucleic acid complexes using RoseTTAFoldNA. *Nat. Methods*, **21**, 117–121.

Basu,S. and Wallner,B. (2016) DockQ: A quality measure for protein-protein docking models. *PLoS One*, **11**, e0161879.

Bernard,C. *et al.* (2024) State-of-the-RNArt: benchmarking current methods for RNA 3D structure prediction. *NAR Genom. Bioinform.*, **6**, lqae048.

Damase,T.R. *et al.* (2021) The limitless future of RNA therapeutics. *Front. Bioeng. Biotechnol.*, **9**, 628137.

Discovery,C. *et al.* (2024) Chai-1: Decoding the molecular interactions of life. *bioRxiv*.

Kagaya,Y. *et al.* (2025) NuFold: end-to-end approach for RNA tertiary structure prediction with

flexible nucleobase center representation. *Nat. Commun.*, **16**, 881.
Kwon,D. (2025) RNA function follows form - why is it so hard to predict? *Nature*, **639**, 1106–1108.
Liu,L. *et al.* (2024) Technical report of HelixFold3 for biomolecular structure prediction. *arXiv [q-bio.BM]*.
Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
Ma,H. *et al.* (2022) Cryo-EM advances in RNA structure determination. *Signal Transduct. Target. Ther.*, **7**, 58.
Mirabello,C. and Wallner,B. (2024) DockQ v2: improved automatic quality measure for protein multimers, nucleic acids, and small molecules. *Bioinformatics*, **40**.
Mirdita,M. *et al.* (2022) ColabFold: making protein folding accessible to all. *Nat. Methods*, **19**, 679–682.
Shen,T. *et al.* (2024) Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nat. Methods*, **21**, 2287–2298.
Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
Wang,W. *et al.* (2023) trRosettaRNA: automated prediction of RNA 3D structure with transformer network. *Nat. Commun.*, **14**, 7266.
Wohlwend,J. *et al.* (2024) Boltz-1 democratizing biomolecular interaction modeling. *bioRxivorg*.
Zhang,C. *et al.* (2023) RMSA: A sequence search and alignment algorithm to improve RNA structure modeling. *J. Mol. Biol.*, **435**, 167904.
Zhang,J. *et al.* (2022) Advances and opportunities in RNA structure experimental determination and computational modeling. *Nat. Methods*, **19**, 1193–1207.
Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*, **33**, 2302–2309.

Tables

Table 1. A brief description of the most important characteristics of benchmarked methods for single-chain RNA and RNA-RNA/Protein prediction.

METHOD	DESCRIPTION	RNA
AlphaFold 3	A deep learning method consisting of a pairformer for residue interaction modeling and a diffusion module for denoising raw atomic coordinates, removing global rotational/translation constraints to enhance prediction flexibility and accuracy.	Single, complex
Boltz-1	AlphaFold 3-inspired architecture with modified order of operations in the MSAModule and transformer layer.	Single, complex
Chai-1	AlphaFold 3-inspired architecture incorporating residue-level embeddings from a large language model for proteins, also introducing experimental constraint features like pocket, contact, and docking constraints. Trained without RNA MSAs.	Single, complex
HelixFold3	AlphaFold 3-inspired architecture using PaddlePaddle library allowing for faster model training.	Single, complex
NuFold	A deep learning method built upon AlphaFold architecture design to predict RNA structures by incorporating nucleic acid sequences, secondary structure inputs, and RNA-specific inter-base angles, distances, and atom positions.	Single
RhoFold+	A deep learning model for RNA structure prediction that utilizes an RNA language model, alongside several deep learning modules, including an IPA module for modeling 3D positions.	Single
RoseTTAFoldNA	A three-track architectured model that simultaneously updates sequence (1D), residue-pair (2D), and structural (3D) representations to predict macromolecular structures..	Single, complex
trRosettaRNA	A deep learning model consisting of MSA and pair representations fed into a transformer network (RNAformer) predicting 1D and 2D geometries, refined in the final 3D structure folding step.	Single

Figures

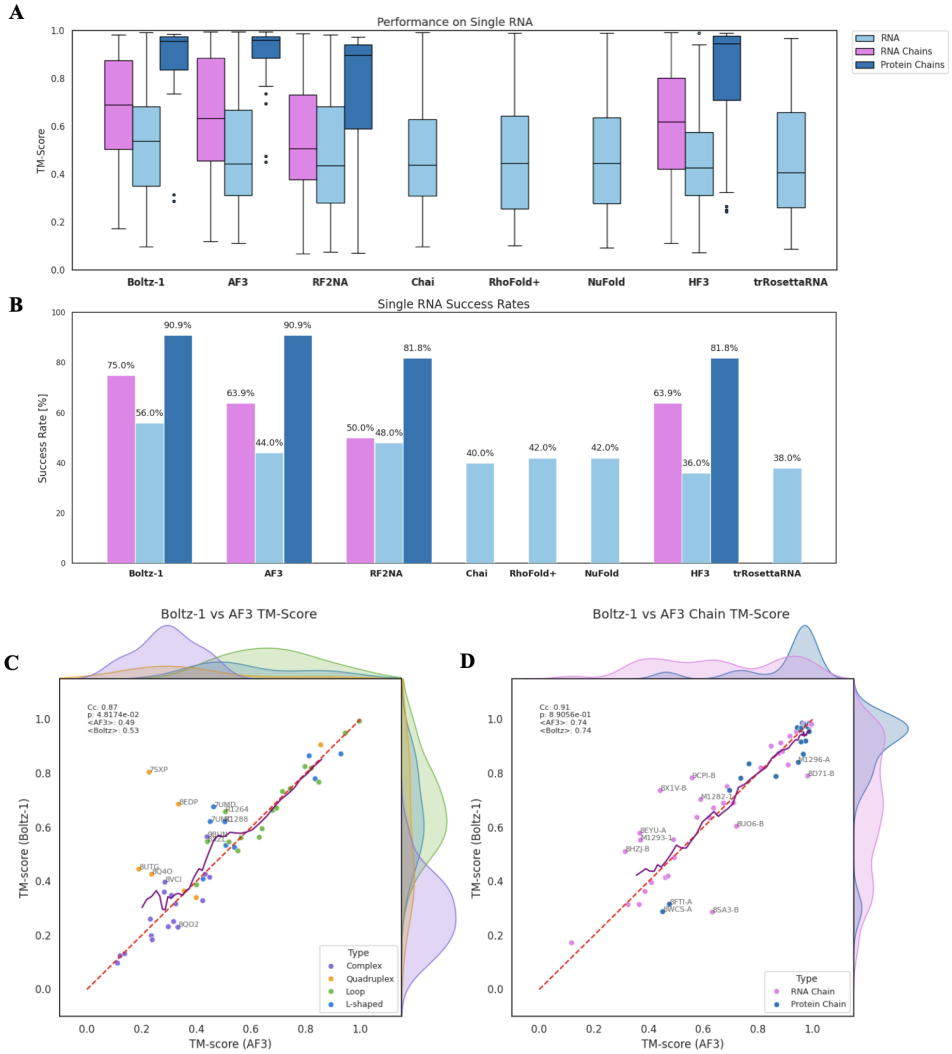
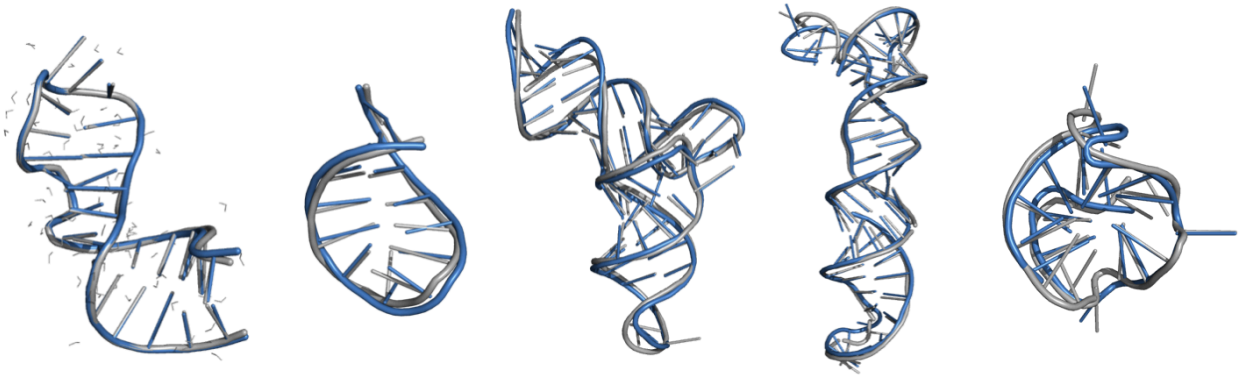


Figure 1. Performance comparison for different RNA prediction methods on single-chain RNA. (A) The performance of RNA prediction models was evaluated using the TM-score. The distribution of TM-scores for each method is presented as a box plot, with the median marked by a line and the interquartile ranges shown. The methods are arranged in descending order of the mean TM-score. Different colors correspond to different datasets: single-chain RNA (light blue), RNA chains derived from complexes (pink), and protein chains derived from complexes (dark blue). (B) Success rates achieved by all methods are defined as the fraction of acceptable models with a minimum TM score of 0.5. (C) Pairplot showing correlation between Boltz-1 (y-axis) and AlphaFold3 (x-axis) TM-scores for single-chain RNA. The labeled points show PDB IDs of predicted models whose difference in TM-score between the two methods is greater than 10%. Paired t-test reported no significant difference between the two methods (Cc = 0.87, P = $4.82*10^{-2}$). (D) Pairplot showing correlation between Boltz-1 (y-axis) and AlphaFold3 (x-axis) TM-scores for single-chain RNA and protein chains (Cc = 0.91, P = $9.3*10^{-1}$).



AlphaFold 3 | 7UCR	NuFold | 8CLR	AlphaFold 3 | 8VT5	Boltz-1 | 8WA4	RF2NA | R1264
TM-Score 0.996	TM-Score 0.950	TM-Score 0.929	TM-Score 0.905	TM-Score 0.905
RMSD 0.31	RMSD 1.19	RMSD 1.43	RMSD 1.78	RMSD 0.47

Figure 2. Top five predicted monomeric RNA structures, ranked by model quality. Superposed predicted RNA is shown in blue, while the ground truth structure is colored in grey. TM scores and RMSD are reported for each model, along with the name of the prediction method that achieved the highest score. Models are displayed in order from best to worst based on TM-score.

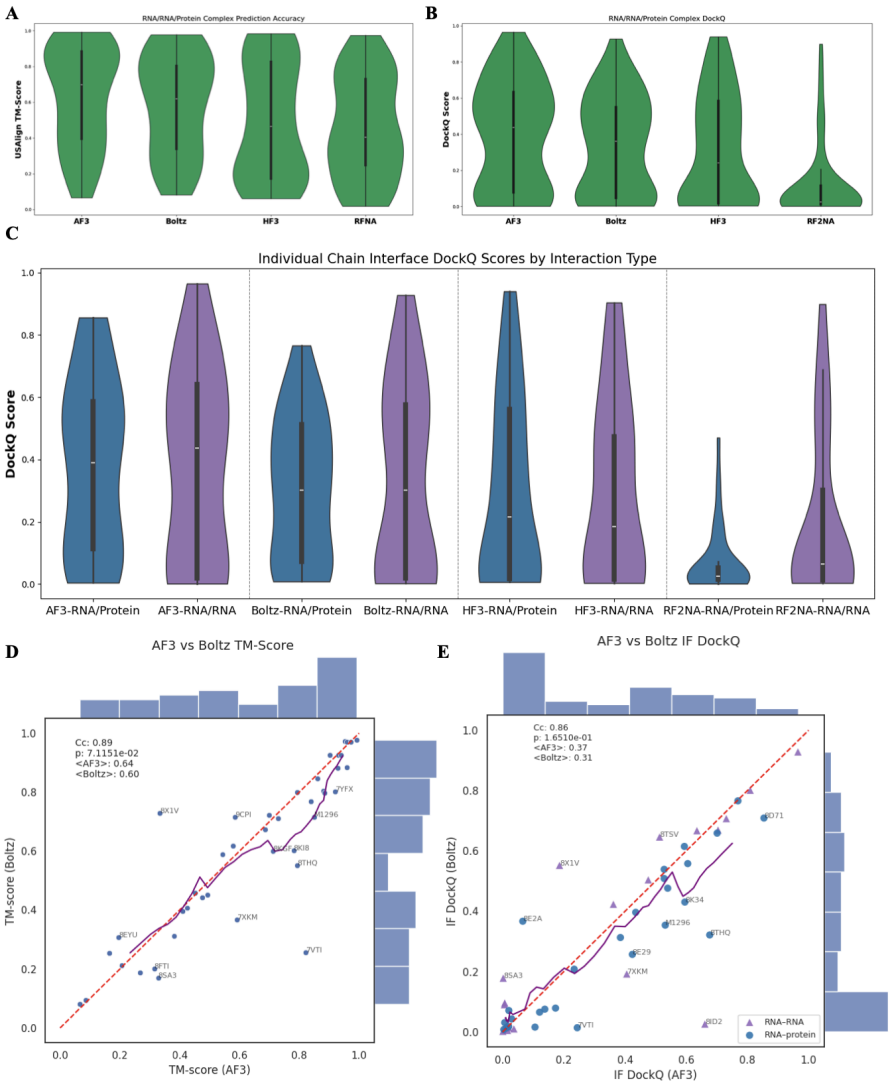


Figure 3. Performance comparison for different RNA prediction methods on RNA complexes. (A) Performance of RNA complex prediction models was evaluated using USAlign TM-score. Each method corresponds to a separate violin plot. The horizontal line inside each violin plot represents the median, while the thin bar in the centre represents the interquartile range. Violin plots are shown in a descending order of TM-score in both cases. (B) Performance of RNA complex prediction models evaluated using DockQ score, following the same layout as in TM-score analysis. (C) DockQ score for individual pairs of chains (Interface DockQ). DockQ scores were calculated separately for RNA-protein (blue) and RNA-RNA (purple) complexes and shown as violin plots arranged in descending order based on the RNA-protein interface DockQ scores. (D) Scatter plot showing correlation between AlphaFold3 and Boltz-1 TM-score for RNA complexes (Cc = 0.89, P = $7.1*10^{-3}$). A rolling average calculated between scores of the two methods is shown as a purple line. Bar plots on each side show the distribution of scores for each model separately (AlphaFold3 on the x-axis, Boltz on the y-axis). E) Scatter plot showing correlation between AlphaFold3 and Boltz-1 interface DockQ score for RNA complexes (Cc = 0.86, P = $1.7*10^{-1}$).

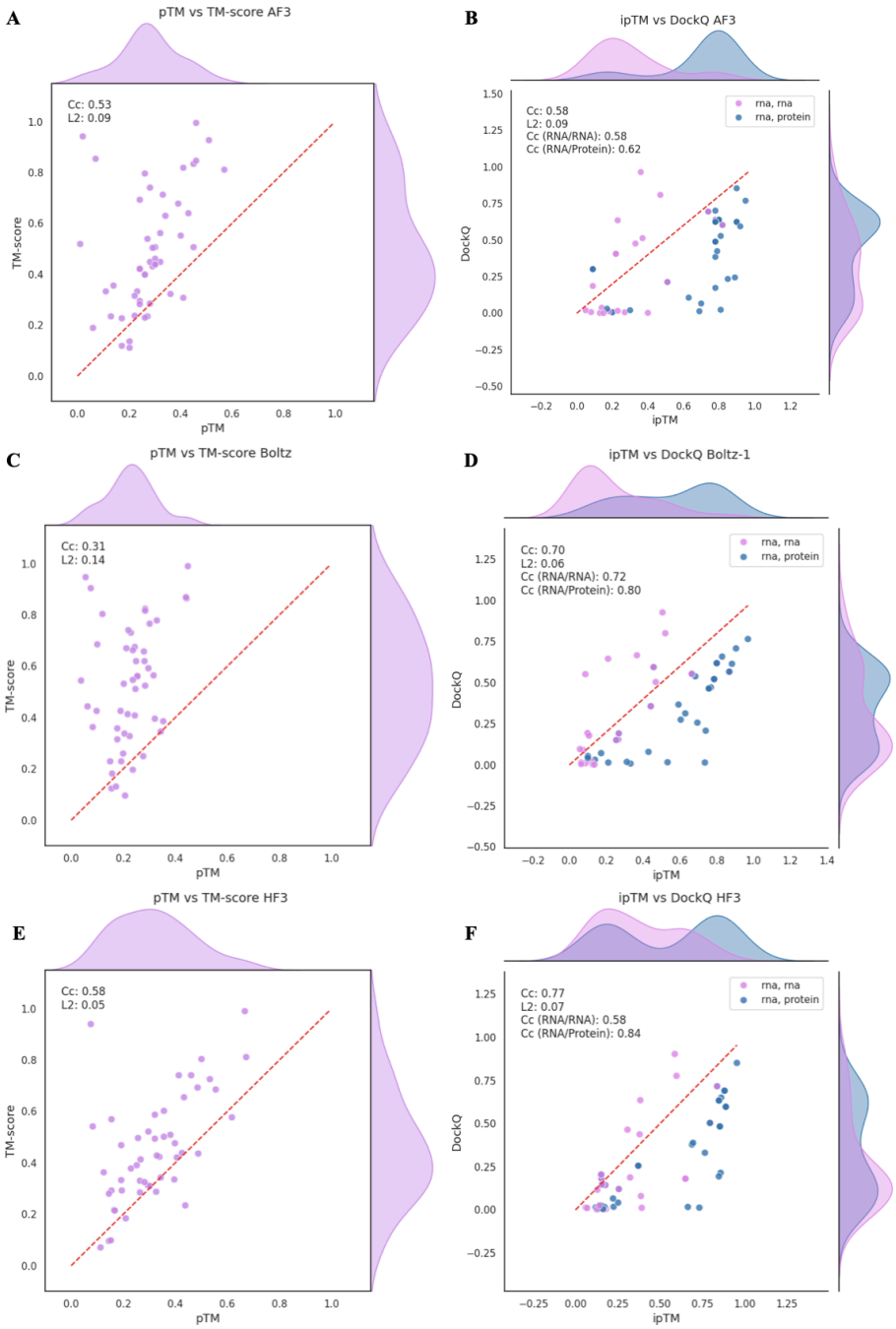


Figure 4. Evaluation of pTM/ipTM confidence scores. (A, C, E) Scatter plot between TM-score and pTM score for AlphaFold3/Boltz-1 and HelixFold3 models for single-chain models. (B, D, F) Scatter plot between DockQ score and ipTM score for RNA complexes, for AlphaFold3, Boltz-1, and HelixFold3. RNA-RNA complexes are colored in pink, and RNA-protein complexes in blue.

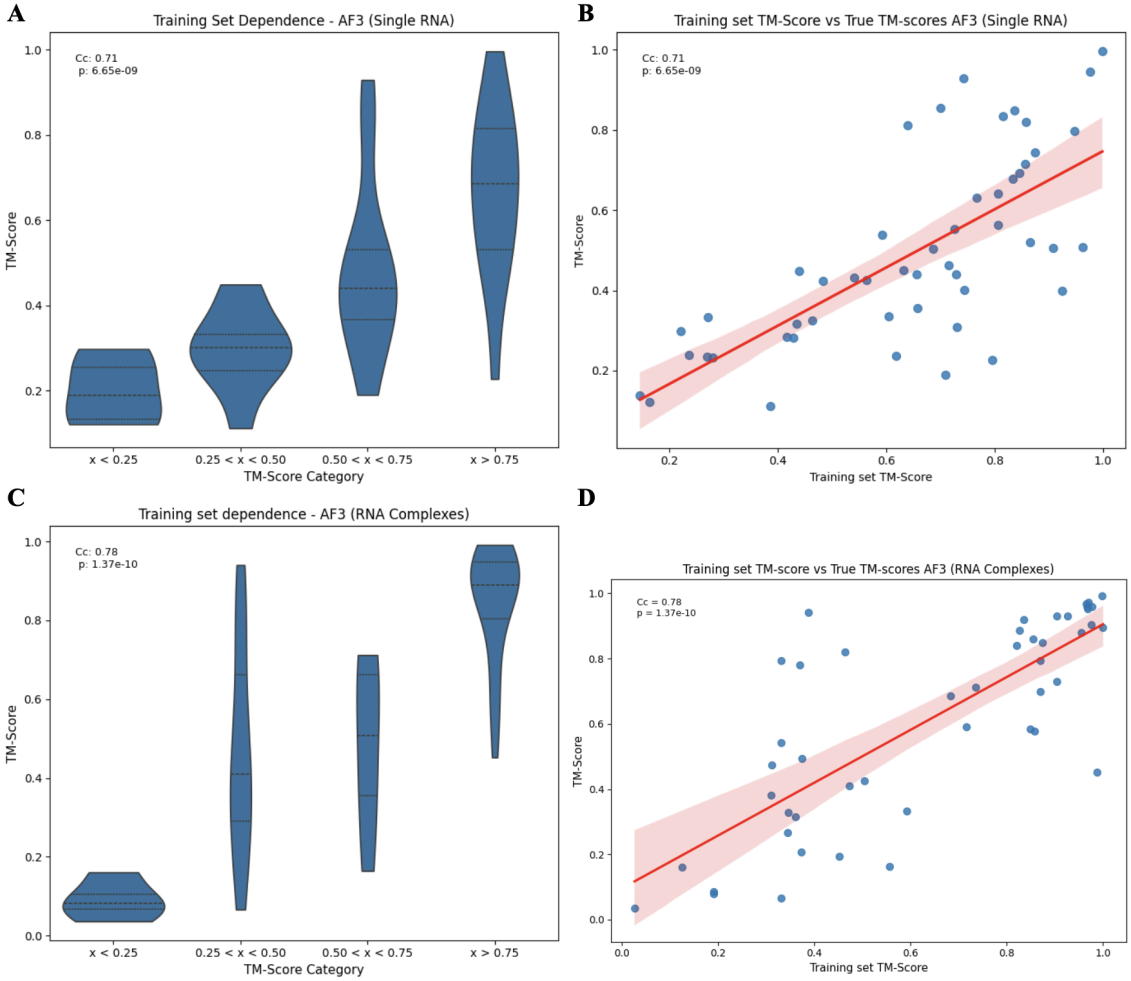


Figure 5. Similarity to the training set. (A) TM-score of benchmarked single RNA models dependent on the training set. TM-scores calculated between structures from the training set (training set TM-score) and benchmarked single RNA structures are divided into four categories. (B) Correlation between TM-scores for benchmarked structures (y-axis) and training set TM-score (x-axis), in the case of AlphaFold3 (Cc = 0.71, P = $6.7*10^{-9}$). (C) TM-score of benchmarked RNA complexes dependent on the training set. TM scores calculated between structures from the training set (training set TM score) and benchmarked RNA complexes are divided into four categories (TM score categories). (D) Correlation between TM-scores for benchmarked structures (y-axis) and training set TM-score (x-axis), in the case of AlphaFold3 (Cc = 0.78, P = $1.4*10^{-1}$).

Supplementary Material for

Limits of deep-learning-based RNA prediction methods

Marko Ludaic, Arne Elofsson*

Supplementary Tables

Table S1. TM-score for the whole benchmarked dataset of single-chain RNA molecules across all tested deep learning methods for single-chain RNA prediction. The most accurate models across all methods for a particular PDB ID are highlighted in bold.

PDB	NuFold	AF3	Chai	RF2NA	HF3	Boltz	RhoFold+	trRosettaRNA
7Q48	0.287	0.356	0.279	0.287	0.289	0.364	0.286	0.265
7SHX	0.243	0.325	0.255	0.247	0.338	0.317	0.233	0.299
7SXP	0.229	0.227	0.233	0.244	0.287	**0.804**	0.230	0.225
7UCR	0.990	**0.996**	0.993	0.983	0.992	0.993	0.990	0.969
7UGA	0.596	0.630	0.543	0.623	**0.695**	0.563	0.596	0.599
7UMC	0.452	0.451	0.454	**0.676**	0.438	0.621	0.453	0.459
7UMD	0.493	0.463	0.567	**0.697**	0.439	0.676	0.489	0.486
7UME	0.824	0.798	0.843	0.847	0.742	0.825	0.828	**0.867**
7V06	0.569	0.679	**0.815**	0.687	0.728	0.663	0.582	0.698
7WIA	0.651	0.563	0.644	**0.702**	0.422	0.561	0.652	0.555
8BM6	0.520	**0.641**	0.532	0.602	0.380	0.595	0.529	0.480
8BTZ	0.299	0.236	**0.470**	0.222	0.235	0.199	0.272	0.259
8BWT	0.716	0.744	0.791	0.721	0.686	0.743	0.710	0.661
8CLR	0.950	0.945	0.938	0.941	0.942	0.949	0.950	0.762
8CQ1	0.804	0.820	0.781	0.794	0.814	0.819	0.806	0.785
8EDP	0.251	0.335	0.221	0.365	0.429	**0.686**	0.213	0.167
8FCS	0.818	**0.848**	0.782	0.800	0.524	0.767	0.817	0.733
8I44	0.535	0.520	0.524	0.665	0.542	0.545	0.575	0.656
8ITS	0.399	0.401	**0.431**	0.425	0.426	0.387	0.395	0.413
8JHP	0.462	0.694	0.815	0.724	0.605	0.671	0.625	0.584
8Q4O	0.297	0.237	0.346	0.347	0.344	0.426	0.296	0.259
8QO2	0.298	**0.333**	0.195	0.250	0.217	0.230	0.278	0.267
8QO4	0.386	0.449	0.266	0.368	0.332	0.415	0.382	0.299
8TNS	0.449	0.400	0.416	0.297	0.365	0.340	0.462	0.452

8UPT	0.808	0.835	0.706	0.615	0.311	0.780	0.806	0.729
8UTG	0.274	0.190	0.425	0.328	**0.469**	0.445	0.225	0.188
8V1H	0.762	0.813	0.712	0.620	0.806	0.865	0.761	0.758
8VCI	0.330	0.285	0.348	0.334	0.392	**0.397**	0.329	0.375
8VPV	0.757	0.508	0.481	0.571	0.335	0.533	0.764	0.682
8VT5	0.765	**0.929**	0.734	0.770	0.589	0.872	0.766	0.822
8WA4	0.295	0.855	0.312	0.279	0.570	0.905	0.260	0.282
8XZL	0.506	0.441	0.461	**0.652**	0.414	0.547	0.498	0.470
9BUN	0.539	0.439	**0.586**	0.519	0.578	0.565	0.500	0.568
9EOW	0.726	0.715	0.644	0.698	0.743	0.733	0.725	**0.811**
R1205	0.436	0.433	0.395	0.449	**0.512**	0.426	0.413	0.313
R1209	0.354	**0.423**	0.352	0.358	0.315	0.329	0.342	0.339
R1211	0.533	0.552	0.512	0.400	**0.656**	0.513	0.539	0.403
R1212	0.238	0.317	**0.350**	0.247	0.282	0.251	0.251	0.313
R1241	0.135	0.309	**0.524**	0.091	0.497	0.348	0.136	0.138
R1242	0.161	0.232	0.224	0.160	0.186	**0.260**	0.142	0.160
R1248	0.206	**0.239**	0.154	0.165	0.215	0.184	0.207	0.201
R1252	0.106	0.121	0.097	0.100	0.096	**0.124**	0.109	0.105
R1256	0.254	0.283	0.313	0.302	0.325	**0.360**	0.254	0.259
R1264	0.839	0.507	0.308	**0.905**	0.503	0.658	0.836	0.882
R1264	0.839	0.507	0.308	**0.905**	0.503	0.658	0.836	0.882
R1281	0.100	0.112	0.128	0.076	0.073	0.097	0.107	0.087
R1285	0.168	**0.298**	0.217	0.155	0.294	0.232	0.184	0.203
R1286	0.093	**0.138**	0.105	0.120	0.100	0.132	0.102	0.093
R1288	0.573	0.505	0.445	0.588	0.293	**0.621**	0.564	0.317
R1293	0.373	0.540	0.371	0.309	0.479	0.526	0.402	0.548
R1296	0.443	0.425	0.385	**0.551**	0.496	0.409	0.442	0.384

Table S2. Mean TM-scores and success rates for each single-chain RNA prediction. Monomeric RNA represents an original benchmark database of single-chain RNAs, while RNA (and protein) chains are chains derived from RNA-RNA and RNA-protein complexes.

	Boltz-1	**AlphaFold3**	**RF2NA**	**Chai1**	**RhoFold+**	**NuFold**	**HelixFold3**	**trRosettaRNA**
Monomeric RNA	0.525/56%	0.491/44%	0.478/48%	0.468/40%	0.466/42%	0.466/42%	0.455/36%	0.453/38%
RNA Chains	0.669/75%	0.651/64%	0.548/50%				0.620/64%	
Protein Chains	0.863/91%	0.887/91%	0.732/82%				0.800/82%	

Table S3. USAlign TM-score for the whole benchmarked dataset of RNA-RNA and RNA-Protein complexes across all tested deep learning methods for complex prediction. The most accurate models across all methods for a particular PDB ID are highlighted in bold.

PDB	RF2NA	Boltz-1	AF3	HF3
7TZR	0.457	**0.459**	0.452	0.417
7UIM	0.392	0.407	**0.425**	0.080
7VSJ	0.577	0.970	0.959	**0.983**
7VTI	0.152	0.256	**0.821**	0.156
7XKM	0.307	0.368	**0.591**	0.391
7YEX	0.571	0.972	0.952	**0.983**
7YEY	0.907	**0.925**	0.903	0.921
7YFQ	0.750	0.805	**0.880**	0.165
7YFX	0.780	0.801	**0.919**	0.912
7YFY	0.760	0.768	**0.839**	0.161
7YGN	0.809	0.798	0.885	**0.907**
8D49	0.099	**0.589**	0.543	0.176
8D4B	0.089	0.452	**0.493**	0.134
8D71	0.770	0.885	**0.959**	0.862
8E29	0.840	0.883	0.929	**0.932**
8E2A	0.680	0.674	**0.685**	0.193
8EYU	0.318	0.308	0.195	**0.388**

8FTI	0.092	0.202	**0.315**	0.149
8HBA	**0.253**	0.213	0.207	0.243
8HZJ	**0.302**	0.255	0.164	0.247
8HZL	0.075	0.095	0.087	**0.137**
8ID2	0.403	0.396	0.410	**0.529**
8K0P	0.037	**0.082**	0.066	0.061
8K34	0.499	0.926	**0.940**	0.881
8KGF	0.033	0.600	**0.712**	0.543
8KI8	0.018	0.602	**0.781**	0.087
8SA3	**0.361**	0.169	0.328	0.244
8SJ7	0.885	0.970	0.972	**0.982**
8SXL	0.974	0.977	**0.991**	0.790
8THQ	0.462	0.551	**0.793**	0.748
8TQX	0.804	0.926	**0.931**	0.916
8TSV	0.292	**0.722**	0.698	0.510
8UO6	0.278	0.311	**0.381**	0.274
8WCS	0.108	0.187	**0.267**	0.123
8WU5	0.707	0.710	**0.730**	0.658
8X1V	0.244	**0.728**	0.333	0.692
9CPI	0.702	**0.716**	0.584	0.636
M1209	0.367	**0.618**	0.578	0.570
M1211	0.411	0.443	**0.475**	0.465
M1282	0.623	**0.800**	0.794	0.192
M1293	0.394	0.846	0.860	**0.893**
M1296	0.831	0.716	**0.850**	0.763

Table S4. DockQ scores for the whole benchmark dataset of RNA-RNA and RNA-Protein complexes across all tested deep learning methods for complex prediction. The most accurate models across all methods for a particular PDB ID are highlighted in bold**.**

PDB	RF2NA	Boltz	AF3	HF3
7TZR	0.009	**0.090**	0.006	0.005
7UIM	0.003	0.008	0.004	**0.015**
7VSJ	0.167	0.079	0.173	**0.215**
7VTI	0.020	0.014	**0.244**	0.017
7XKM	0.229	0.192	**0.406**	0.181
7YEX	0.174	**0.539**	0.528	0.194
7YEY	**0.014**	0.014	0.013	0.013
7YFQ	0.028	0.595	**0.638**	0.182
7YFX	0.009	0.313	**0.385**	0.379
7YFY	0.018	0.554	**0.603**	0.205
7YGN	0.012	0.659	**0.701**	0.331
8D49	0.046	0.208	**0.234**	0.017
8D4B	0.015	0.191	**0.213**	0.144
8D71	0.348	0.709	**0.854**	0.651
8E29	0.283	0.257	**0.424**	0.389
8E2A	0.169	**0.367**	0.066	0.018
8EYU	0.011	**0.096**	0.007	0.080
8FTI	0.013	**0.031**	0.007	0.009
8HBA	0.014	**0.020**	0.013	0.014
8HZJ	0.016	0.010	**0.037**	0.012
8HZL	0.003	0.014	**0.020**	0.011
8ID2	0.077	0.053	**0.301**	0.256
8K0P	**0.035**	0.002	0.001	0.003
8K34	0.026	0.522	**0.624**	0.598
8KGF	0.012	**0.615**	0.594	0.042
8KI8	0.001	0.016	**0.106**	0.066
8SA3	0.108	**0.178**	0.002	0.117
8SJ7	0.470	0.766	0.769	**0.852**
8SXL	0.899	0.928	**0.965**	0.904
8THQ	0.207	0.357	0.695	**0.717**
8TQX	0.689	0.801	**0.809**	0.777

8TSV	0.079	**0.646**	0.513	0.188
8UO6	0.004	0.005	**0.016**	0.003
8WCS	0.013	0.018	0.020	**0.025**
8WU5	**0.524**	0.504	0.476	0.438
8X1V	0.136	**0.552**	0.186	0.465
9CPI	0.653	**0.667**	0.635	0.636
M1209	0.016	0.471	**0.489**	0.484
M1211	0.020	**0.071**	0.023	0.017
M1282	0.017	**0.043**	0.030	0.006
M1293	0.020	0.619	0.639	**0.691**
M1296	0.284	0.465	0.625	**0.634**

Table S5. Mean TM-scores, DockQ scores, and interface DockQ scores for each RNA Complex prediction method. Mean IF DockQ excludes scores for protein-protein interface DockQ and is averaged across RNA-protein and RNA-RNA interface DockQ.

	AlphaFold3	**Boltz-1**	**HelixFold3**	**RF2NA**
TM-score	0.635	0.597	0.502	0.462
DockQ	0.394	0.339	0.309	0.122
Interface DockQ	0.372	0.312	0.281	0.118

Table S6. Mean interface DockQ scores for RNA-protein and RNA-RNA interfaces.

	AlphaFold3	**Boltz-1**	**HelixFold3**	**RF2NA**
RNA-protein	0.382	0.309	0.288	0.072
RNA-RNA	0.355	0.317	0.269	0.190

Supplementary Figures

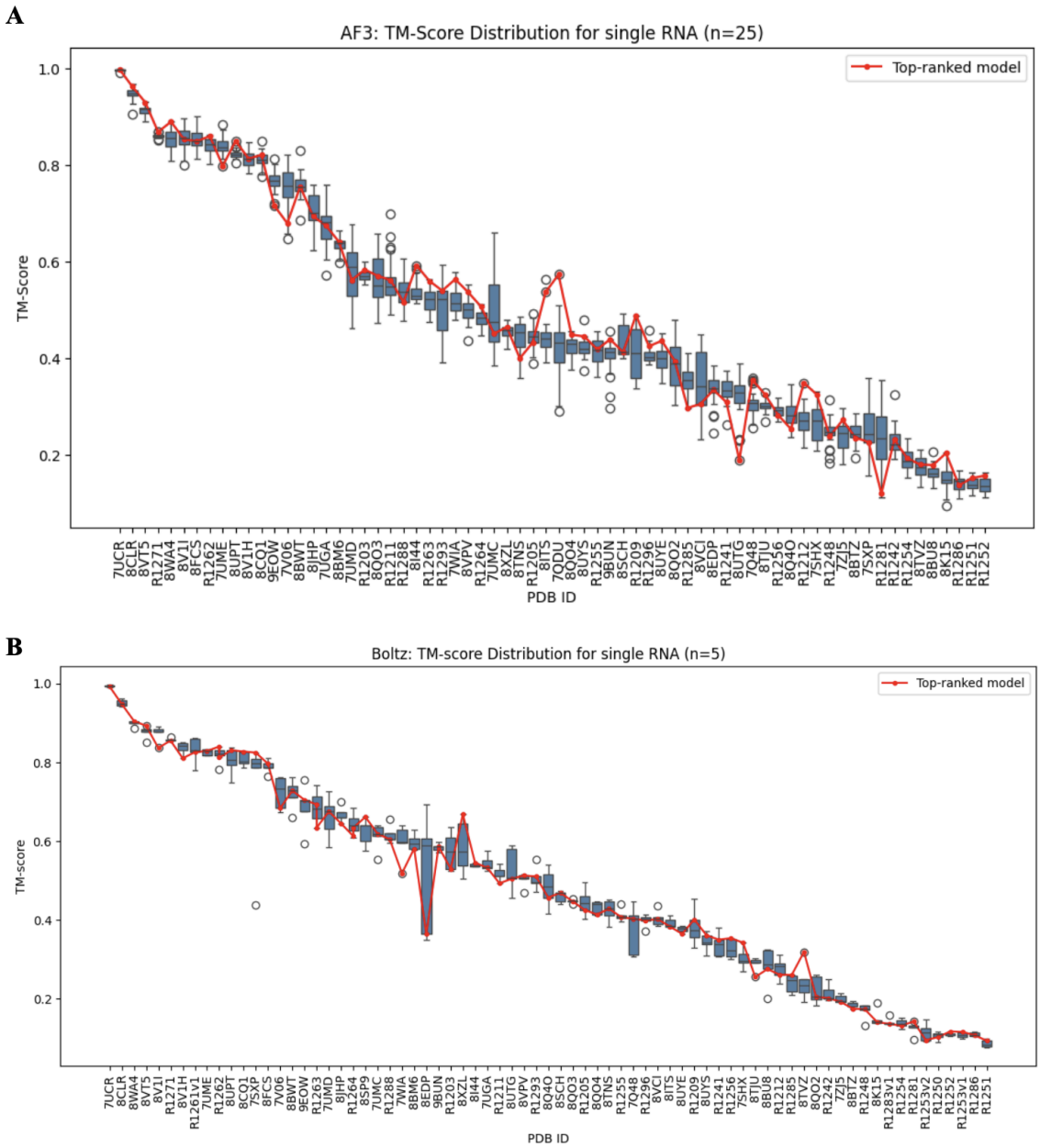


Figure S1. Distribution of TM-scores across *n* generated single-chain RNA models for AlphaFold3 (A) and Boltz-1 (B). TM scores for each ID are shown as box plots, ordered from highest to lowest mean TM score. The red line across the plot connects the top-ranked models of each ID.

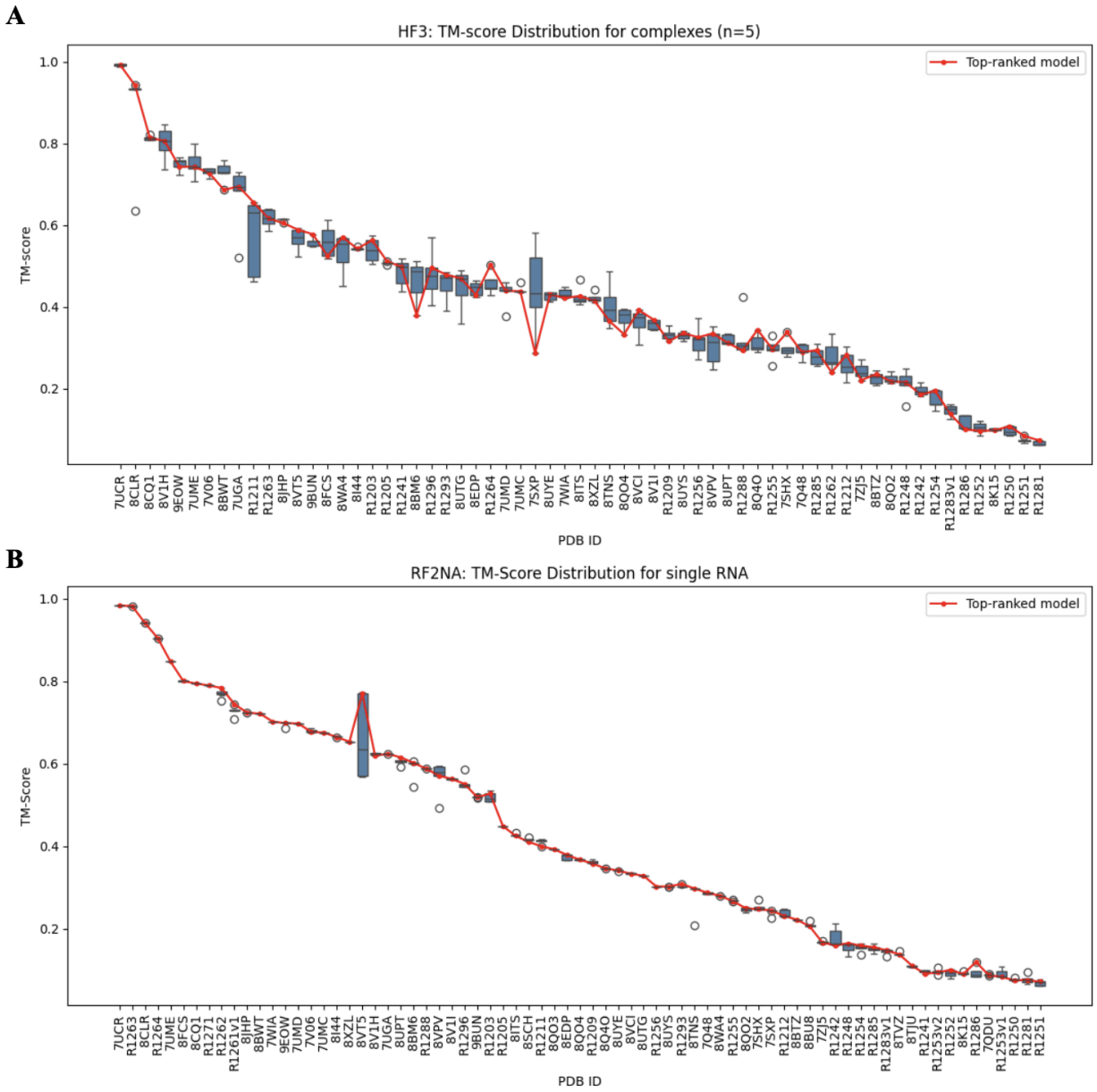


Figure S2. Distribution of TM-scores across n generated single-chain RNA models for HF3 (A) and RF2NA1 (B). For each ID, TM scores are visualized as box plots, ordered from highest to lowest mean TM score. A red line connects the top-ranked model for each ID.

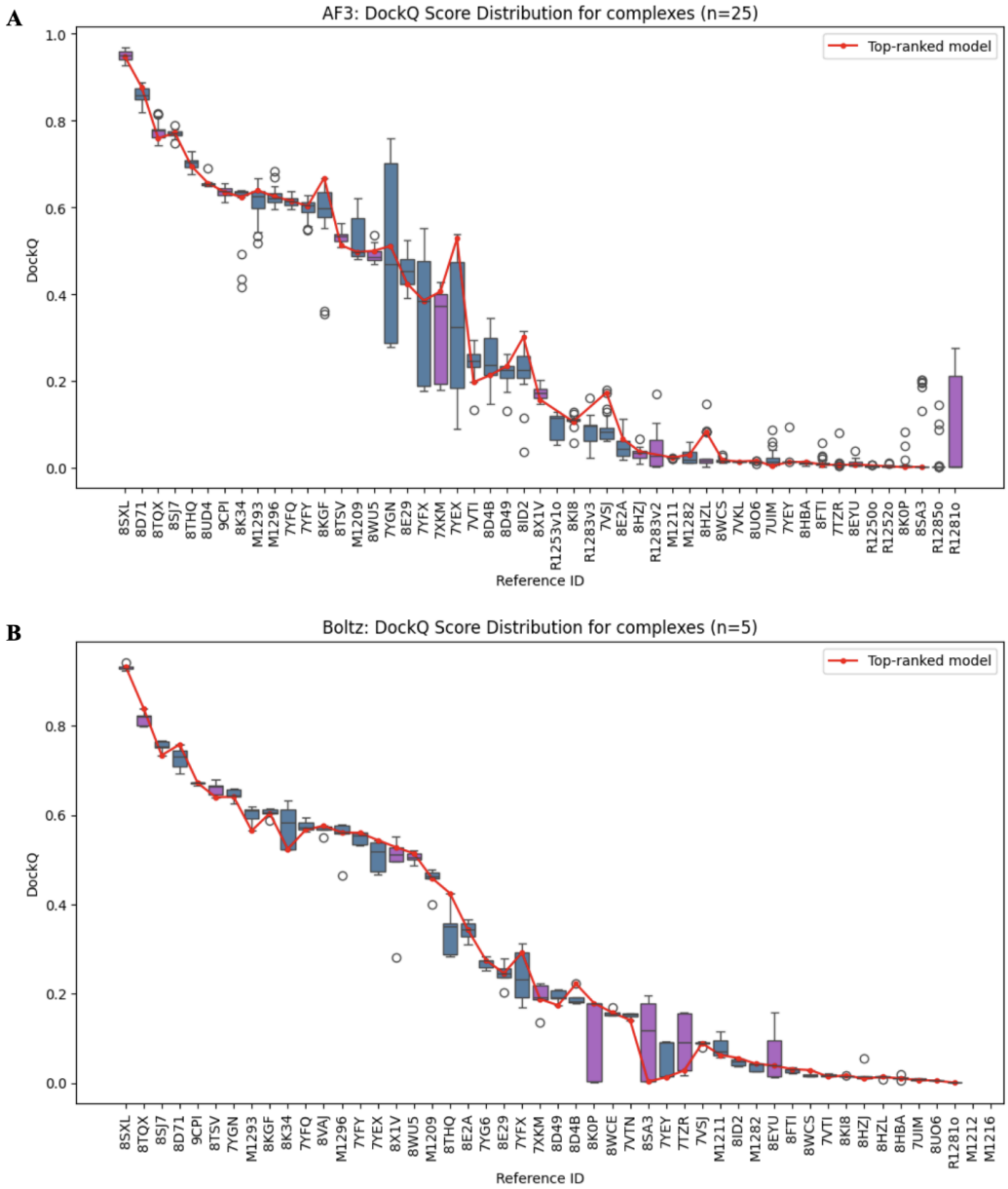


Figure S3. Distribution of TM-scores across *n* generated RNA complex models for HF3 (A) and RF2NA1 (B). For each ID, TM scores are visualized as box plots, ordered from highest to lowest mean TM score. Box plots are colored blue for RNA-protein complexes and purple for RNA-RNA complexes. A red line connects the top-ranked model for each ID.

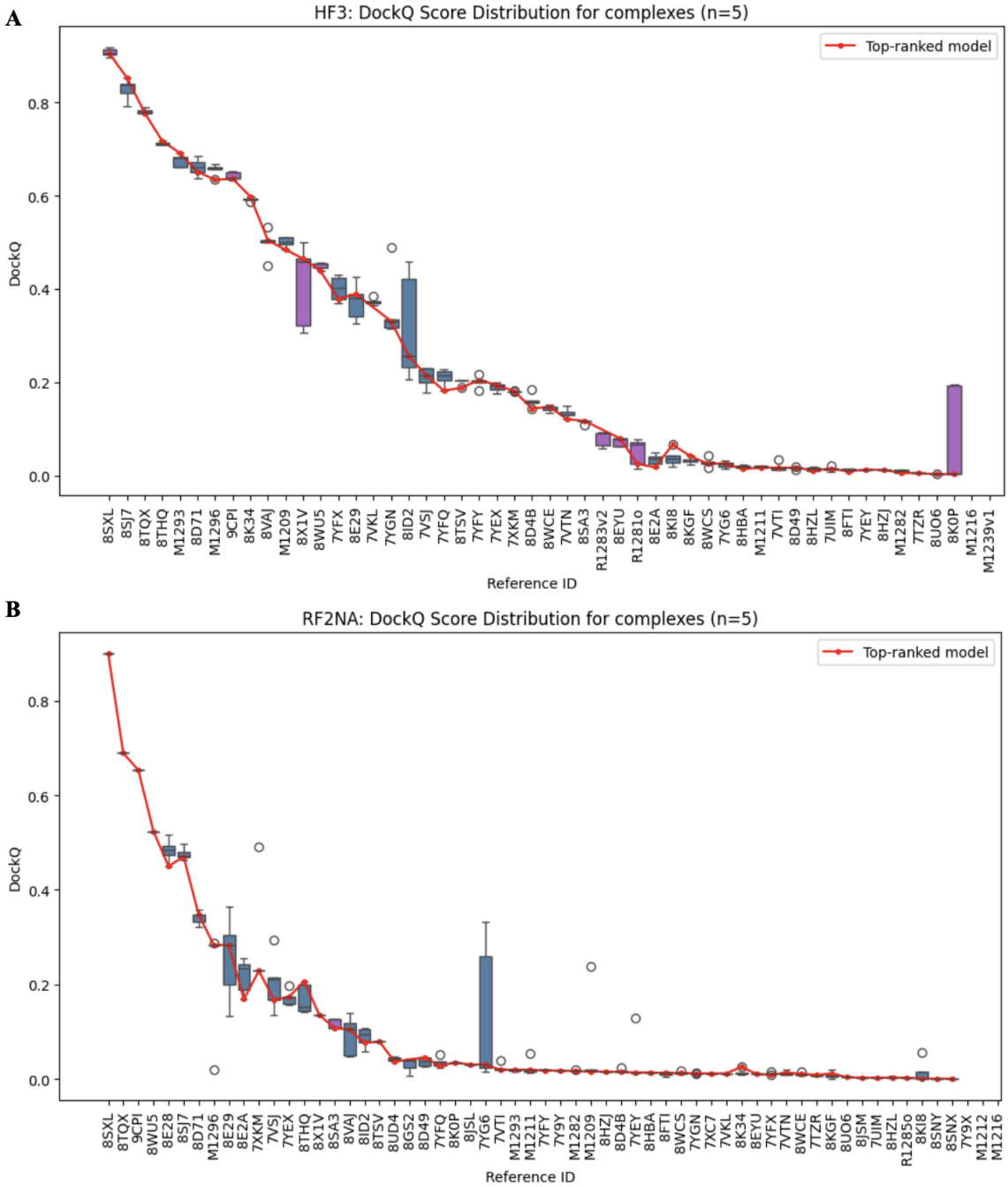


Figure S4. Distribution of TM-scores across *n* generated RNA complex models for HF3 (A) and RF2NA1 (B). For each ID, TM-scores are visualized as box plots, ordered by decreasing mean TM-score. Box plots are colored blue for RNA-protein complexes and purple for RNA-RNA complexes. A red line connects the top-ranked model for each ID.

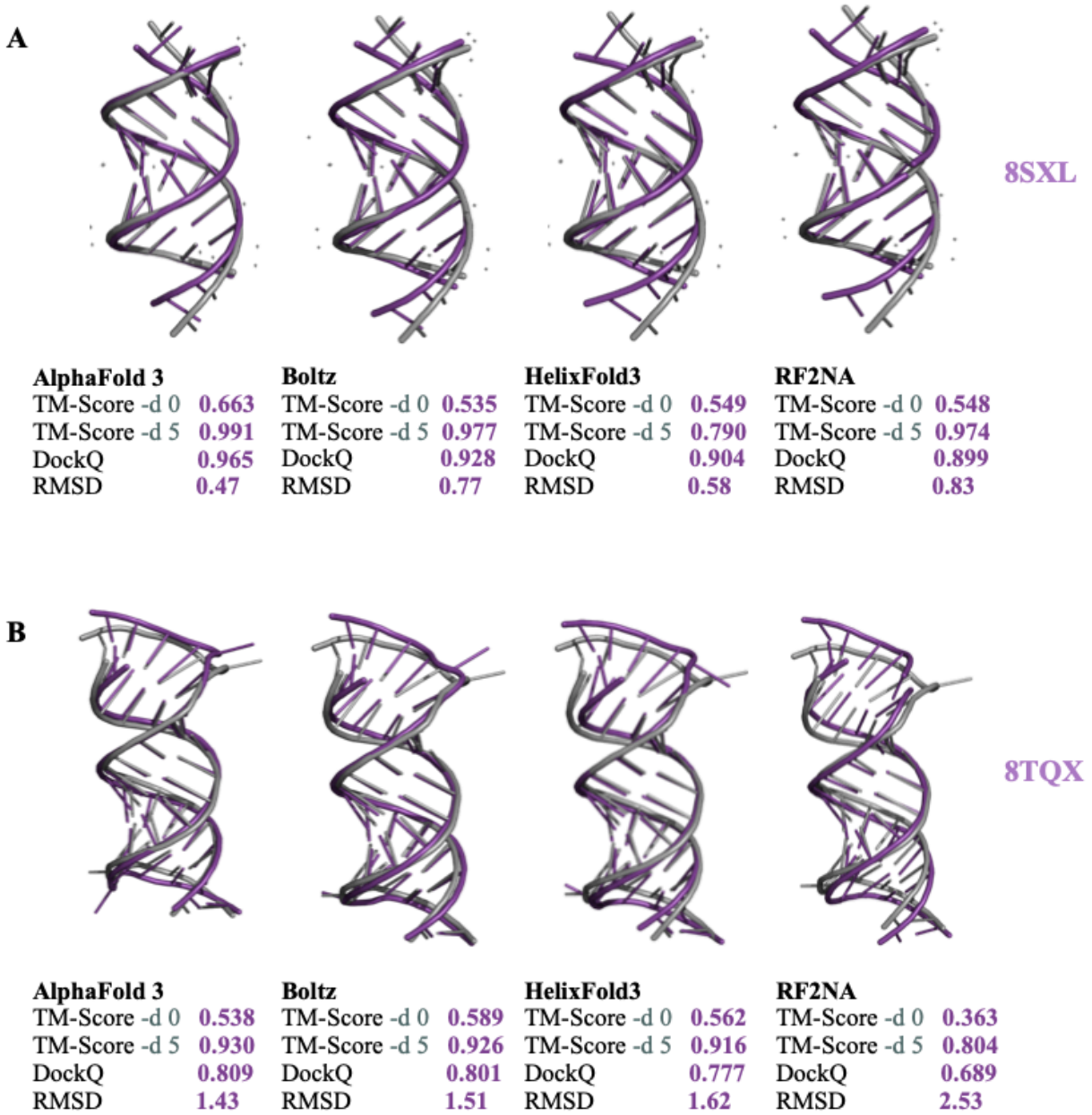


Figure S5. Structural alignment of the highest-scoring structure (8SXL) and another double-helical highly-scoring structure (8TQX). The ground truth is presented in grey, while the modeled structure is given in purple. TM-scores both before and after *-d* parameter correction are shown, along with DockQ score and RMSD.

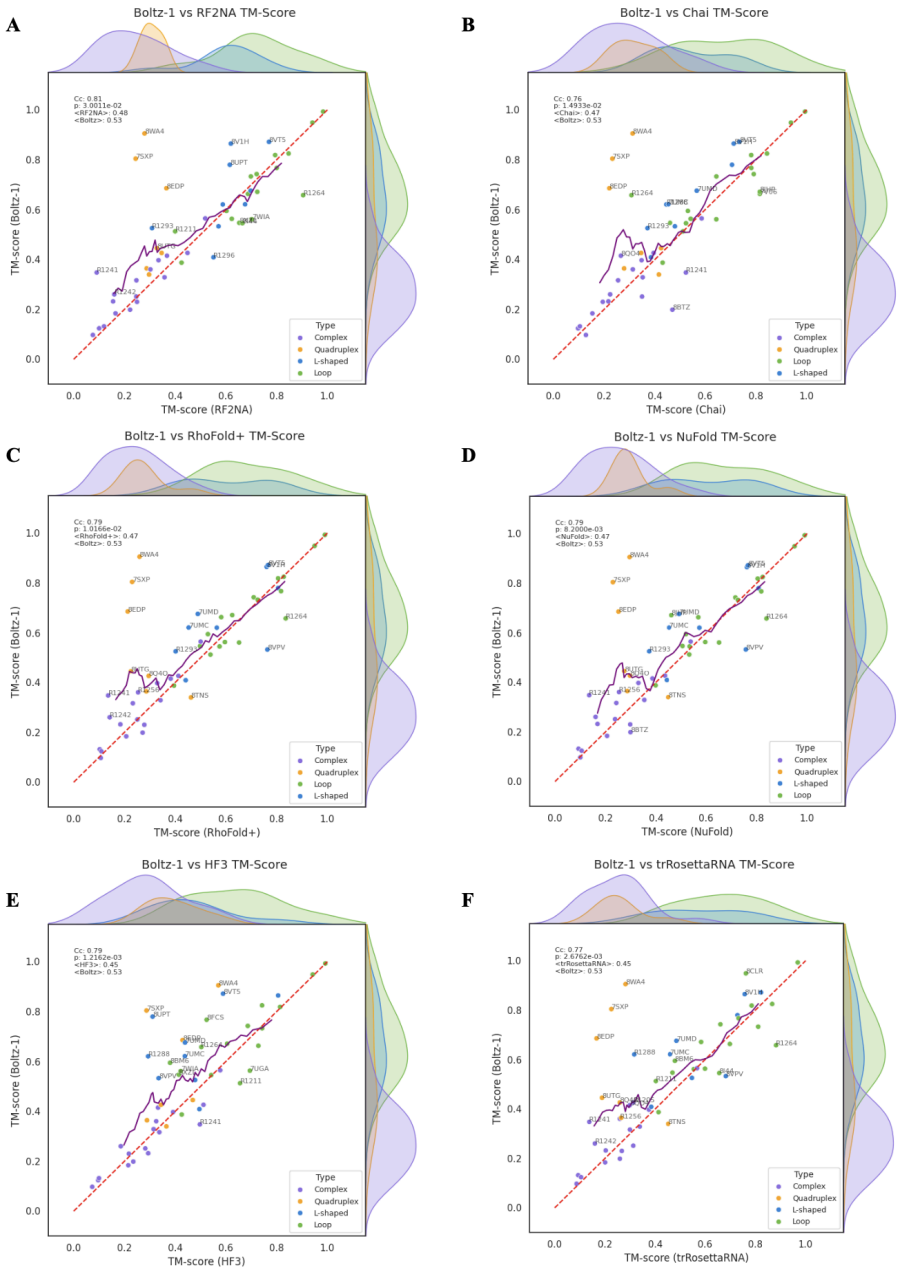


Figure S6. Pair plots showing TM-score correlation between the top-ranked method (Boltz-1) against TM-scores of other methods. Cc denotes the Pearson correlation coefficient, paired T-test p-value by *p*. Significant differences in TM-scores were found between Boltz-1 and Nufold ($P = 8.2*10^{-3}$), HF3 ($P = 1.2*10^{-3}$), and trRosettaRNA ($P = 2.7*10^{-3}$).

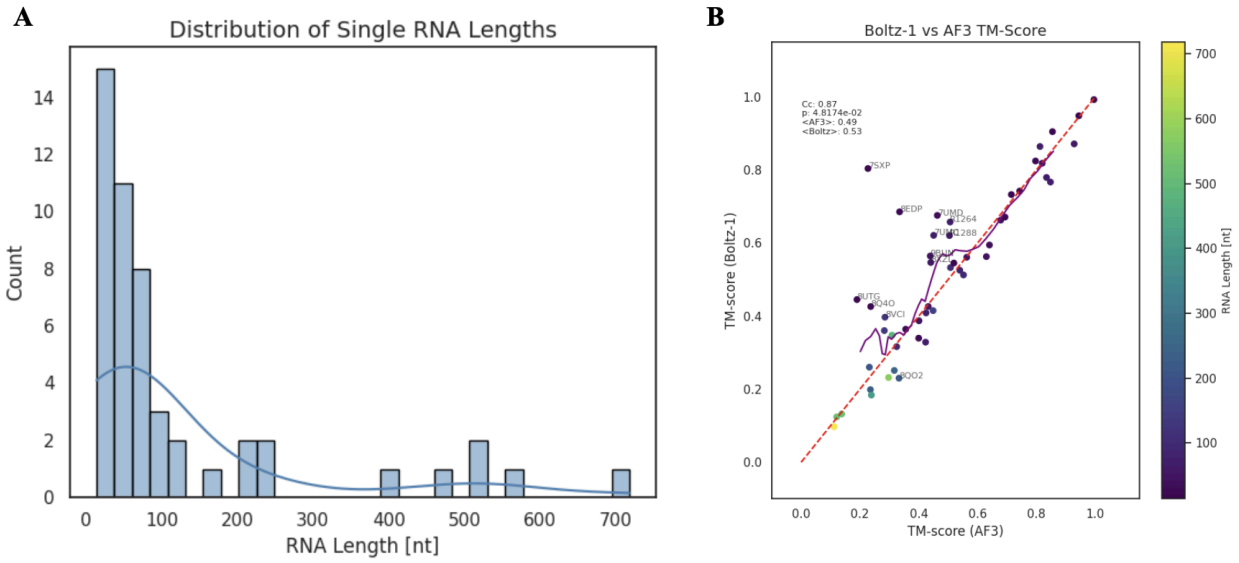


Figure S7. (A) Distribution of single-chain RNA sequence lengths. (B) TM-score scatter plot between Boltz-1 and AlphaFold3 colored by RNA length [nt, nucleotide]. The majority of longer RNAs (> 300 nt) have a TM-score below 0.5.

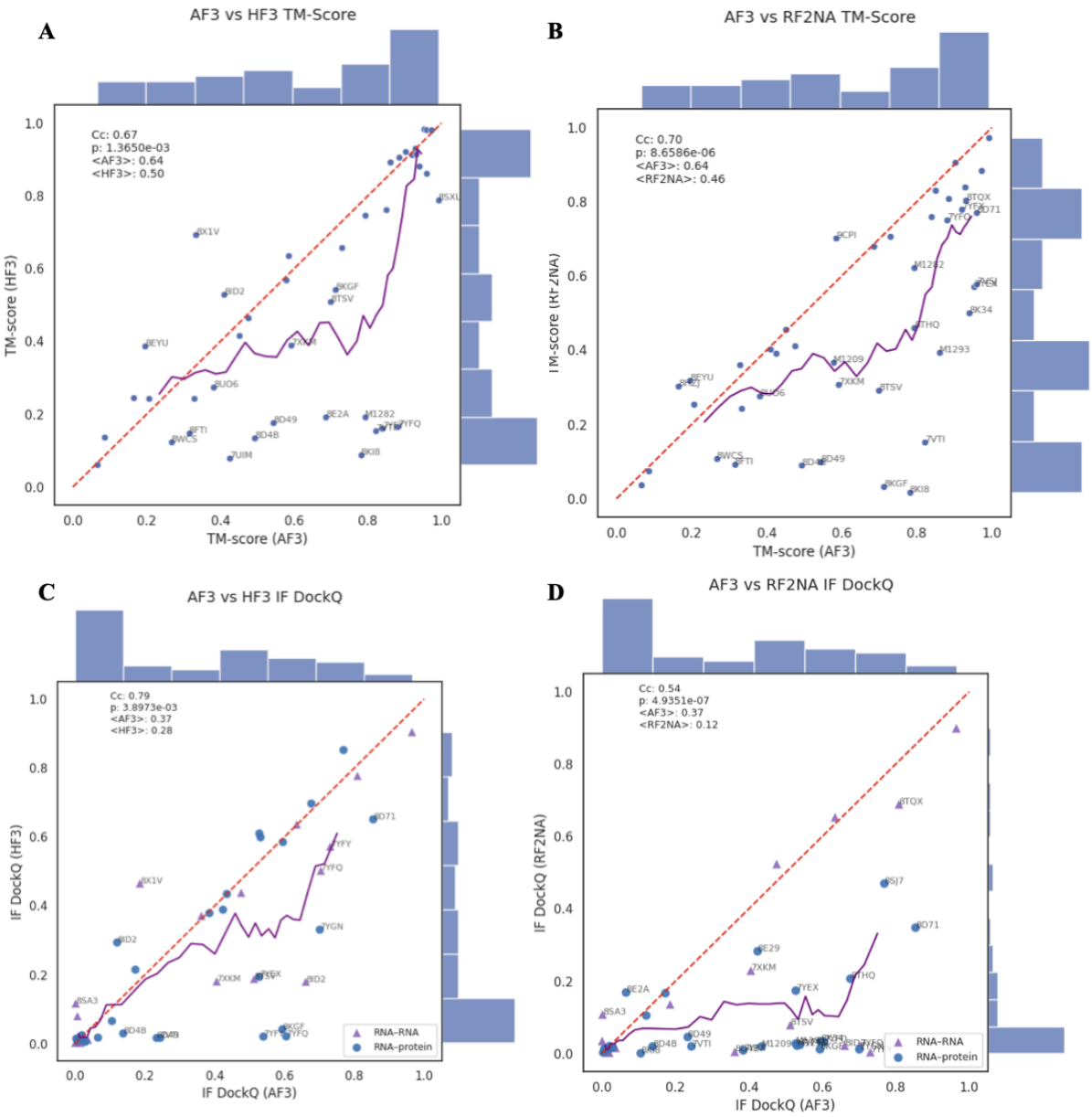


Figure S8. Scatter plots showing (A, B) TM-score and (C, D) Interface DockQ score correlation between the top-ranked method (Boltz-1) against the corresponding scores of other methods. Cc denotes the Pearson correlation coefficient, and the paired T-test p-value by p. Significant differences in TM-scores and interface DockQ scores were found between AlphaFold3 and Boltz-1, and AlphaFold3 and RF2NA, for both TM-score and interface DockQ score.

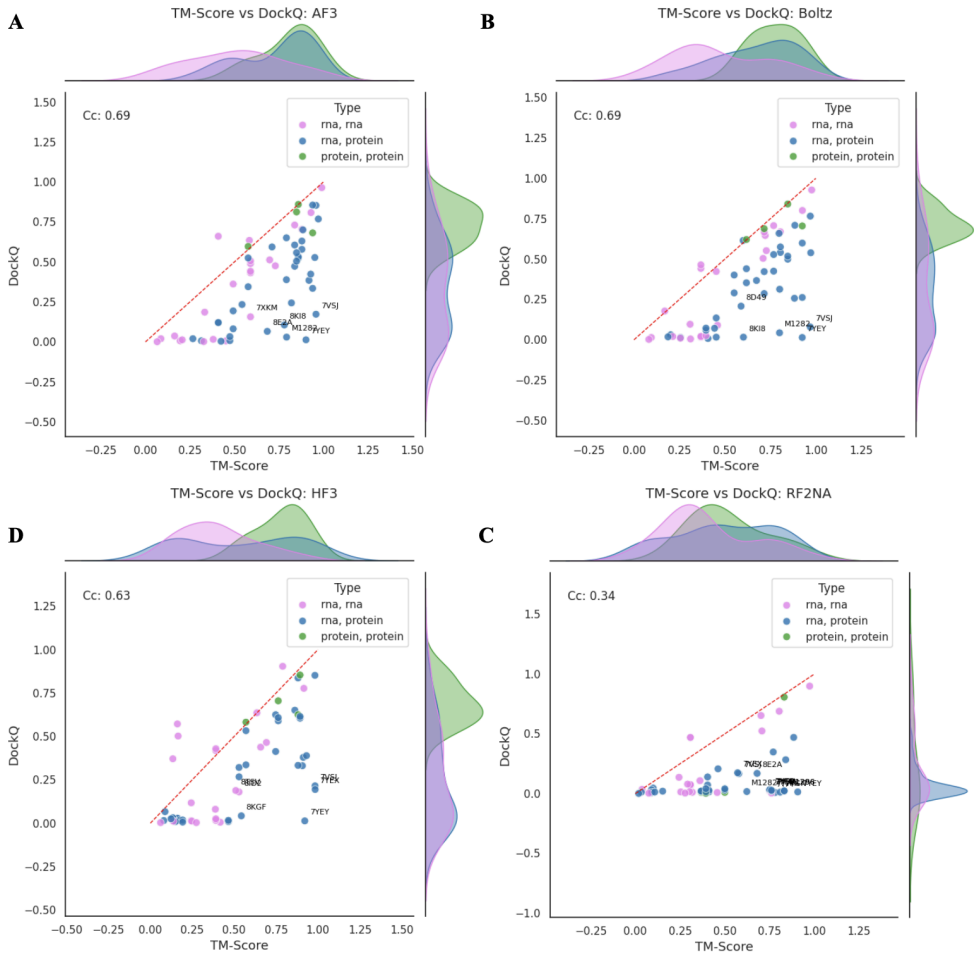


Figure S9. Scatter plots showing the correlation between TM score and DockQ interface. Points are colored based on the interface type (RNA-protein in blue, RNA-RNA in pink, Protein-protein in green). Plots are presented for each RNA complex prediction method: (A) AlphaFold3, (B) Boltz-1, (C) HelixFold3, (D) RF2NA. Labeled points show PDB IDs with high TM scores (> 0.5) but low interface DockQ scores (< 0.23).

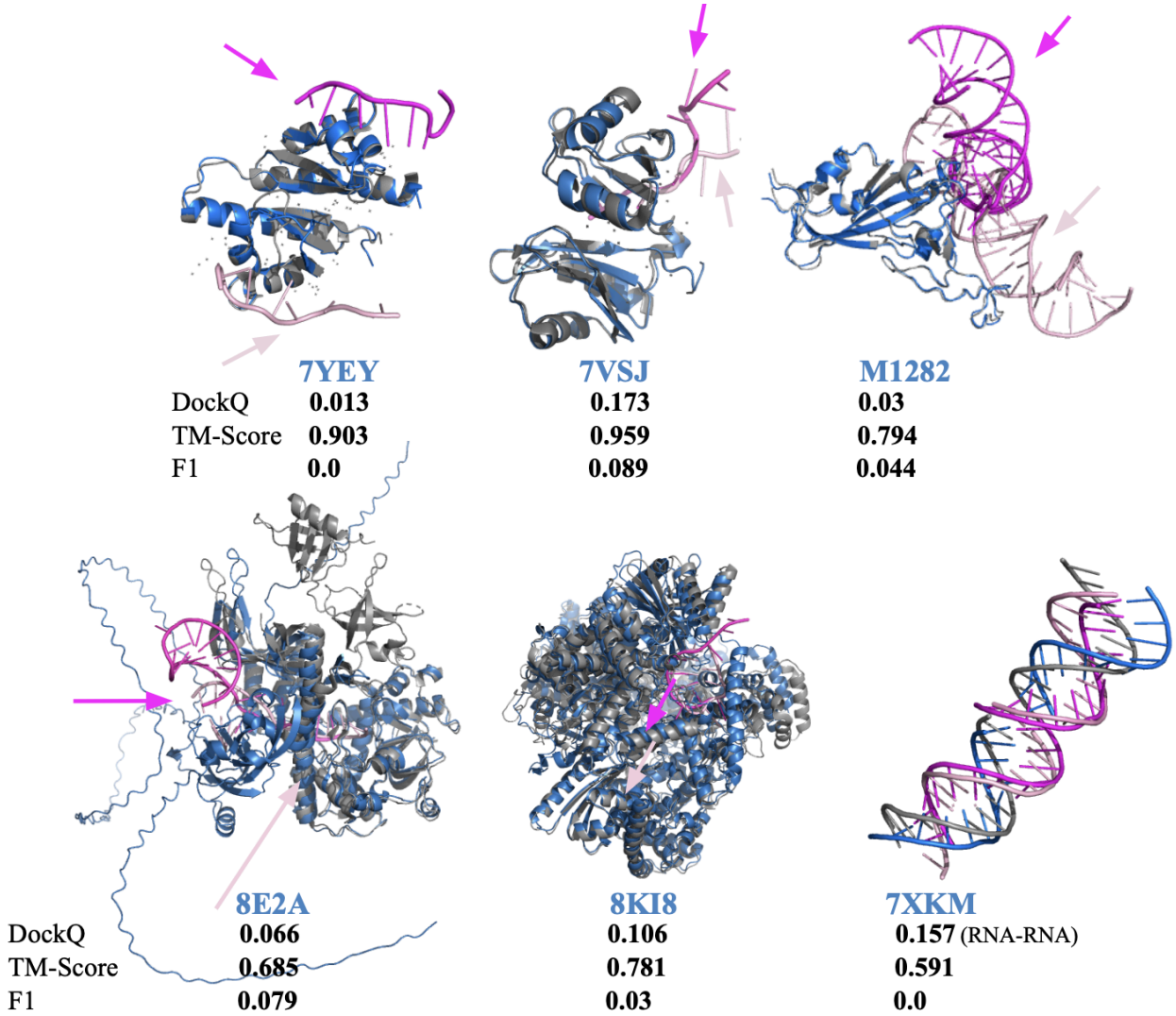


Figure S10. Set of structures with high TM-score (> 0.5) and low (interface) DockQ score (< 0.23). In all 5 RNA-protein cases, the total DockQ score is the same as the interface DockQ, as the whole complex is made of one RNA and one protein chain. In complexes where the protein shows internal symmetry (7YEY, 7VSJ), RNA is misplaced on the opposite side of the protein chain. In other cases, the RNA binding site is incorrectly predicted. In the case of tetrameric 7XKM, the RNA-RNA interface with DockQ 0.157 and F1 0.0 corresponds to the Chain A-Chain C interface (pink) between the two homodimers. In all cases, RNA chains are colored in light pink in ground truth structures and magenta in predicted models.

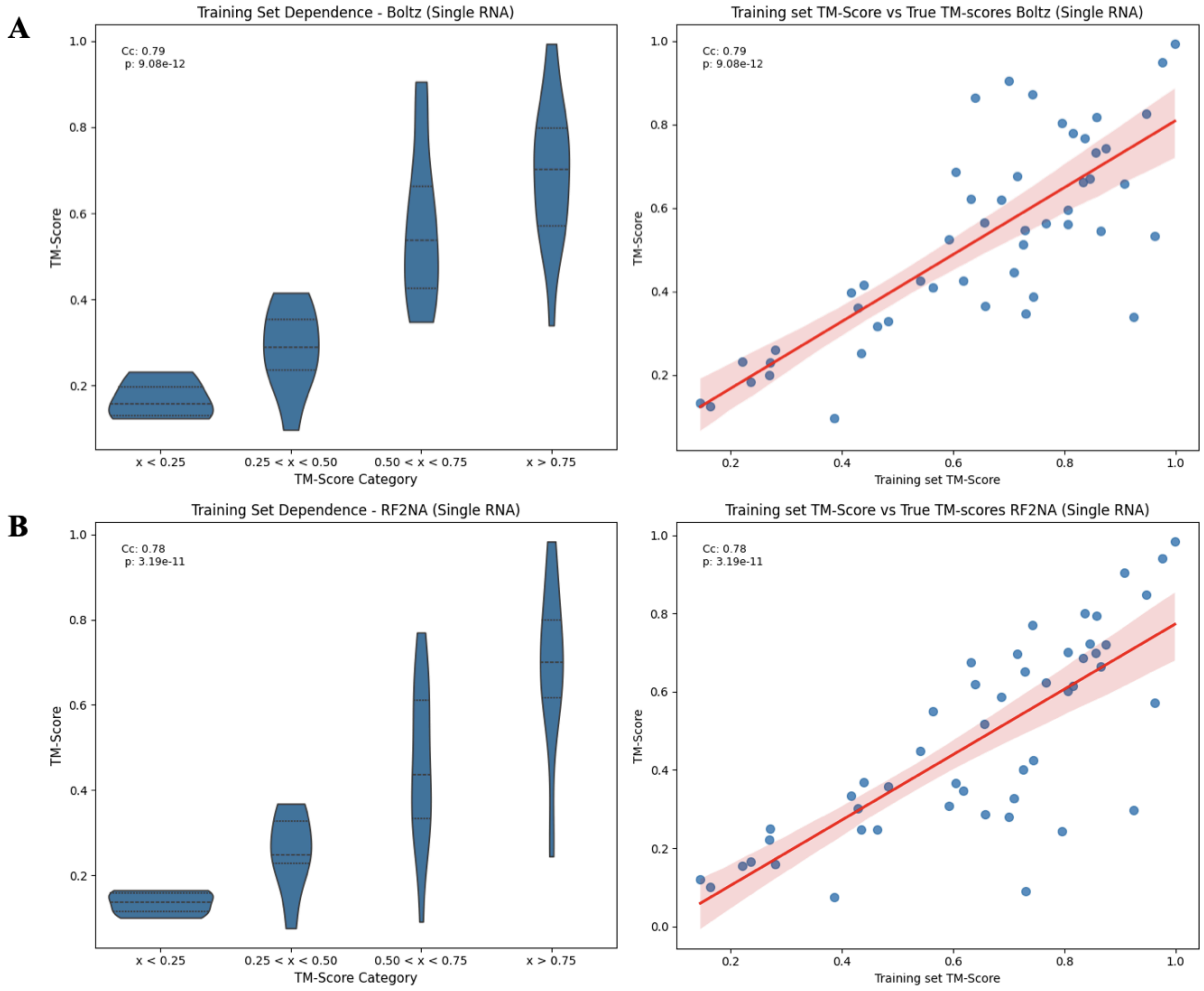


Figure S11. (A) Violin plot showing TM-score training set categories for single RNA and Boltz-1 (left). Correlation plot between TM-scores for the benchmarked structures (y-axis) and training set TM-score (x-axis), in the case of Boltz-1 (right). (B) Violin plot showing TM-score training set categories for single RNA and RF2NA (left). Correlation plot between TM-scores for benchmarked structures (y-axis) and training set TM-score (x-axis), in the case of RF2NA (right).

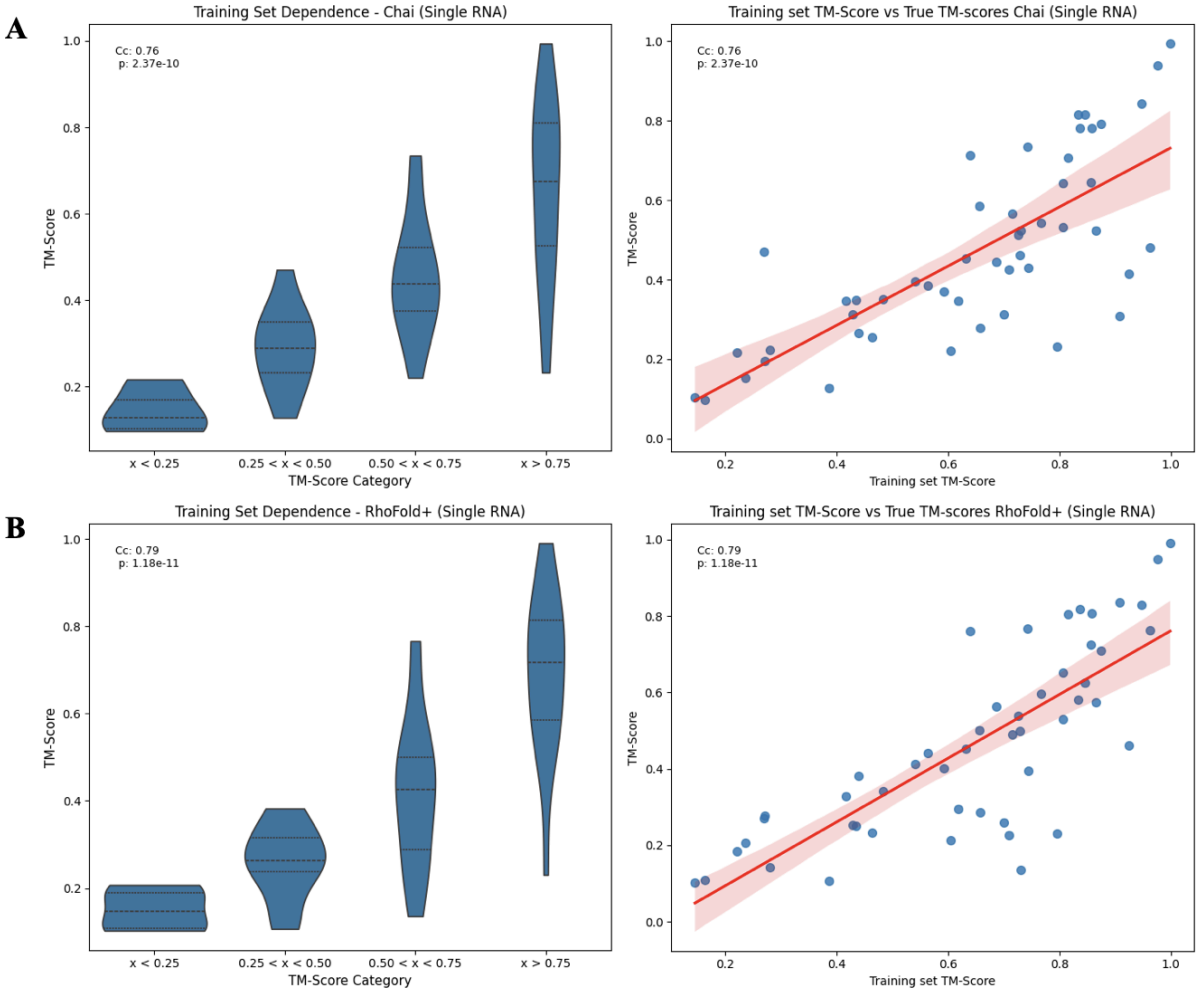


Figure S12. (A) Violin plot showing the TM-score training set categories for single RNA and Chai (left). Correlation plot between TM-scores for benchmarked structures (y-axis) and training set TM-score (x-axis), in the case of Chai (right). (B) Violin plot showing TM-score training set categories for single RNA and RhoFold+ (left). Correlation plot between TM-scores for benchmarked structures (y-axis) and training set TM-score (x-axis), in the case of RhoFold+ (right).

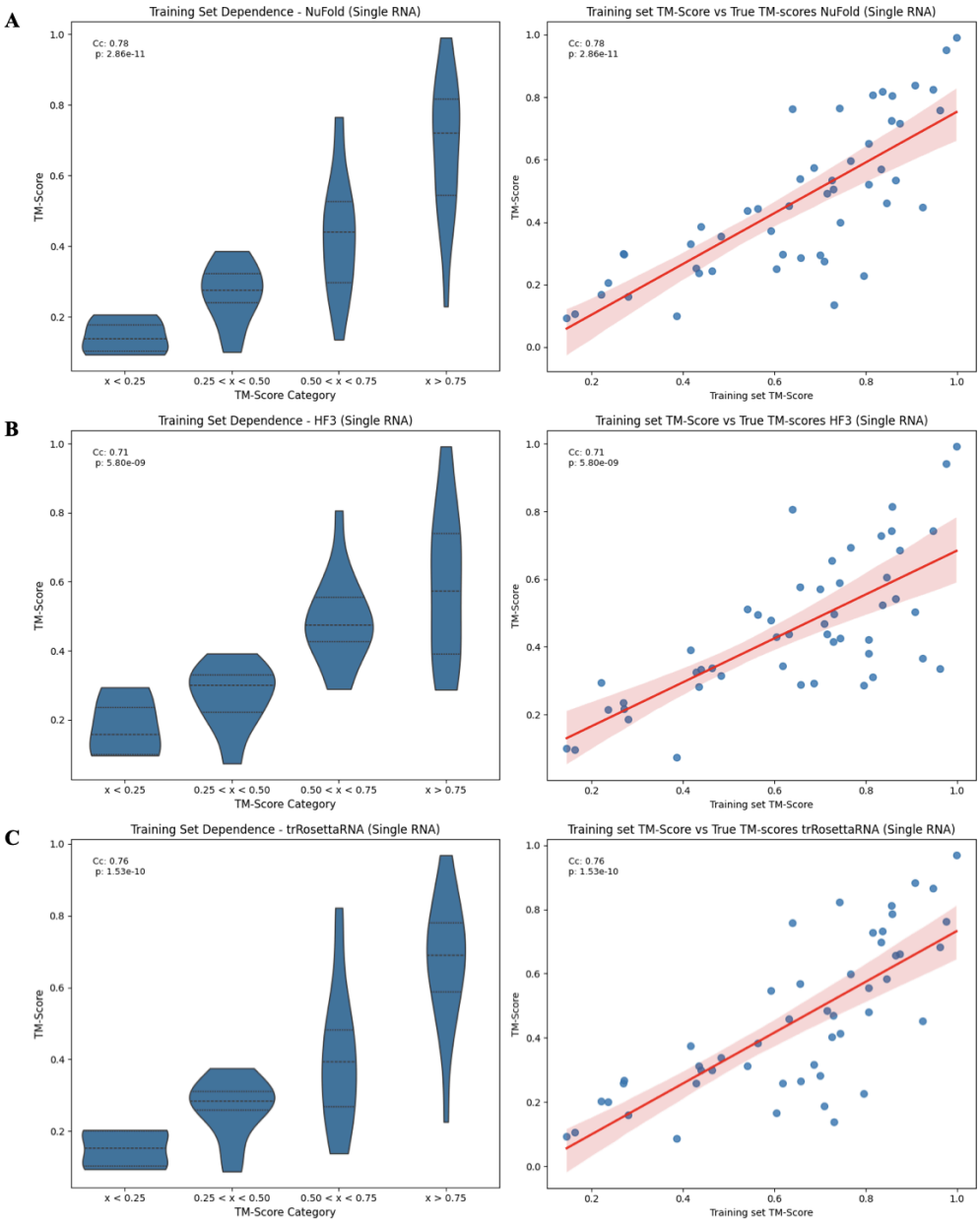


Figure S13. (A) Violin plot showing TM-score training set categories for single RNA and NuFold (left). Correlation plot between TM-scores for benchmarked structures (y-axis) and training set TM-score (x-axis), in the case of NuFold (right). (B) Violin plot showing TM-score training set categories for single RNA and HelixFold3 (left). Correlation plot between TM-scores for benchmarked structures (y-axis) and training set TM-score (x-axis), in the case of HelixFold3 (right). (C) Violin plot showing TM-score training set categories for single RNA and trRosettaRNA (left). Correlation plot between TM-scores for benchmarked structures (y-axis) and training set TM-score (x-axis), in the case of trRosettaRNA (right).

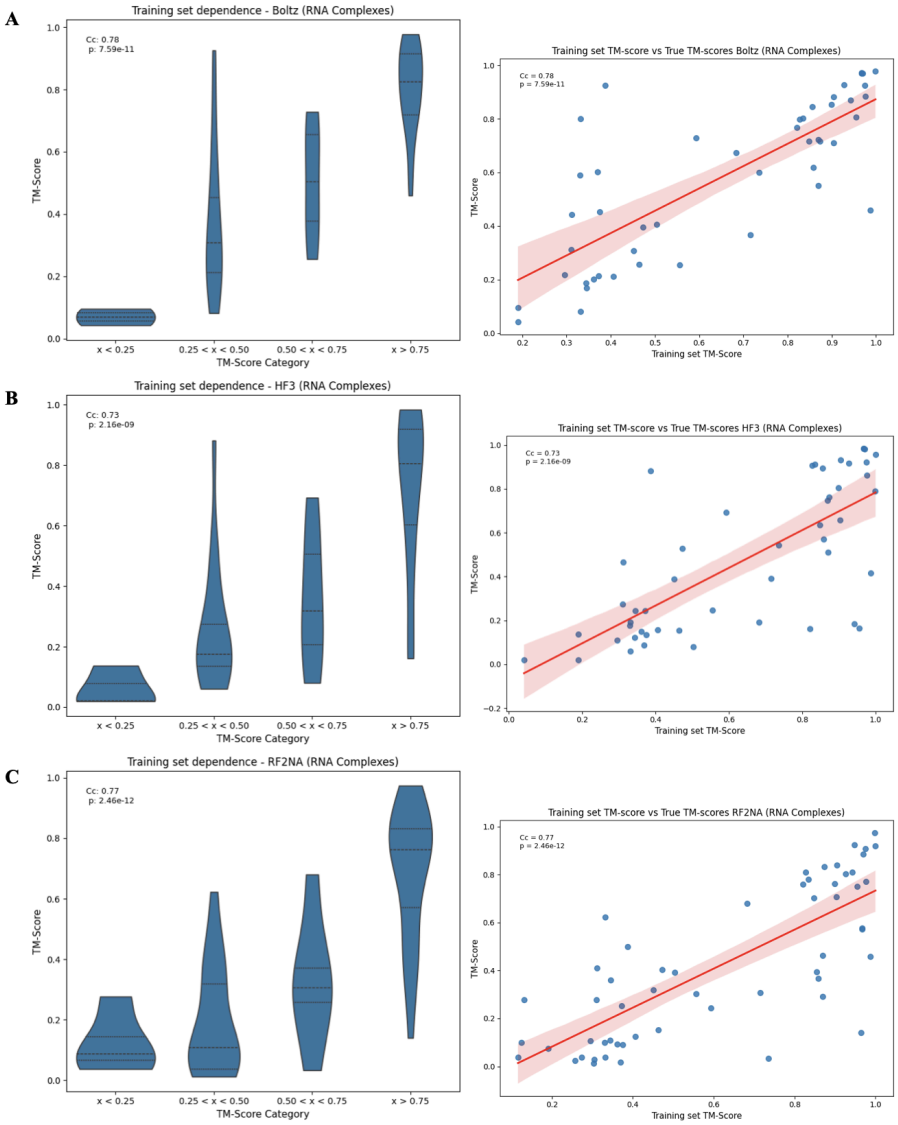


Figure S14. (A) Violin plot showing TM-score training set categories for RNA complexes and Boltz-1 (left). Correlation plot between TM-scores for benchmarked structures (y-axis) and training set TM-score (x-axis), in the case of Boltz-1 (right). (B) Violin plot showing TM-score training set categories for RNA complexes and HelixFold3 (left). Correlation plot between TM-scores for benchmarked structures (y-axis) and training set TM-score (x-axis), in the case of HelixFold3 (right). (C) Violin plot showing TM-score training set categories for RNA complexes and RF2NA (left). Correlation plot between TM-scores for benchmarked structures (y-axis) and training set TM-score (x-axis), in the case of RF2NA (right)