Generalization and Scoring in RNA 3D Structure Prediction: A Benchmarking Study

Ivona Martinović[a], **Tin Vlašić**[b], **Yang Li**[c], **Bryan Hooi**[c], **Yang Zhang**[c], **Mile Šikić**[d]

[a] *National University of Singapore (NUS) & Agency for Science, Technology and Research (A*STAR), Singapore* ivona_martinovic@gis.a-star.edu.sg

[b] *A*STAR, Singapore* tin_vlasic@gis.a-star.edu.sg

[c] *NUS, Singapore* zhang@nus.edu.sg, bhooi@comp.nus.edu.sg, liyangum@nus.edu.sg

[d] *A*STAR, Singapore & University of Zagreb, Croatia* mile_sikic@gis.a-star.edu.sg

* Presenting author

1. Introduction

RNA molecules play essential roles in gene regulation, catalysis, and other biological processes, with their functions tightly linked to their 3D structures [1]. While experimental methods like X-ray crystallography and cryo-EM provide high-resolution structures, they are costly and time-intensive [2], motivating computational approaches for 3D structure prediction. Traditional computational methods, including physics-based and knowledge-based approaches, have been widely used for the structure prediction [3–9], but their accuracy remains limited. The success of AlphaFold models [10, 11] in protein structure prediction demonstrated the potential of deep learning, leading to the development of several deep learning-based models for RNA structure prediction [12–19].

Recent reviews [20–23] and comparative studies [24–28] highlight the growing interest in RNA 3D structure prediction. However, the existing benchmarks often lack systematic dataset design, with some using small or evaluation sets that overlap with or closely resemble training datasets, limiting their ability to assess generalization. Additionally, comparisons among all available deep learning-based models remain incomplete, with AlphaFold 3 often omitted or tested separately.

Our main contribution is twofold. First, we construct GenRNA, a dataset where test RNAs are distinct from training data, and benchmark six RNA structure prediction models: DRfold [12], DeepFoldRNA (DFR) [13], RhoFold [16], RoseTTAFoldNA (RF2NA) [29], trRosettaRNA (trRNA) [18], and AlphaFold 3 (AF3) [19]. Implementations of epRNA [14] and NuFold [15] were not available at the time of this study. We assess the models' ability to generalize to unseen RNA sequences and evaluate whether performance remains within acceptable thresholds. Second, we investigate ensemble-based selection using scoring functions. Although several scoring functions exist [30–32], we focus on ARES [33] and Rosetta score [34], as they are among the most widely used. We assess whether scoring functions can improve structure selection and explore the potential for further refinement in model ranking strategies.

2. Methods

Dataset: GenRNA. We introduce GenRNA, a dataset designed to assess model generalization by ensuring that evaluation RNAs were sequentially distinct from those used for training of each of the six models. Since most models do not disclose their training data, we used 13 January 2023, the validation cutoff for AlphaFold 3, as a universal training cutoff, as this date is after all evaluated models were developed. RNA structures published in the Protein Data Bank (PDB) [35] were clustered at 90% sequence identity. Clusters containing only RNAs deposited after the mentioned date were selected and further filtered based on length, resolution, and completeness. To ensure comparability, we included only RNAs for which all models produced final structures, resulting in 84 sequences. A detailed dataset construction pipeline is provided in Appendix A.

Scoring functions. We evaluated whether ARES and Rosetta score could reliably identify the best prediction among the six models and thereby improve overall structure prediction accuracy. ARES is a deep learning-based model trained on RMSD values, while Rosetta score is an energy-based function. We first analyzed how well each score correlated with RMSD and then evaluated whether choosing the structure identified as the best by ARES or Rosetta score resulted in a lower RMSD compared to relying only on individual model predictions. As part of this analysis, we introduce the optimal function, which always selects the prediction with the lowest RMSD value. This serves as an upper bound on performance, representing the best possible accuracy achievable with existing deep learning models when paired with a perfect scoring function.

3. Results and Discussion

Figure 1 shows all-atom Root Mean Square Deviation (RMSD), a metric that quantifies the structural deviation between predicted and reference RNA structures, calculated using the RNA-Puzzles Toolkit [36]. Median RMSD values range from 9.64 Å for DeepFoldRNA to 16.33 Å for AlphaFold 3, indicating that the current models struggle to accurately predict RNA structure, as RMSD values below 5 Å are generally considered acceptable, and values be-

low 2 Å indicate high accuracy [36]. In comparison, when evaluated on the RNA Puzzles dataset [37–41], which has significant overlap with training data, models achieve much lower RMSD values, with the lowest median 2.65 Å for RoseTTAFoldNA and the highest 7.92 Å for AlphaFold 3 (Appendix B). This stark contrast highlights the models' reliance on training-set similarity and their limited ability to generalize to unseen RNAs.

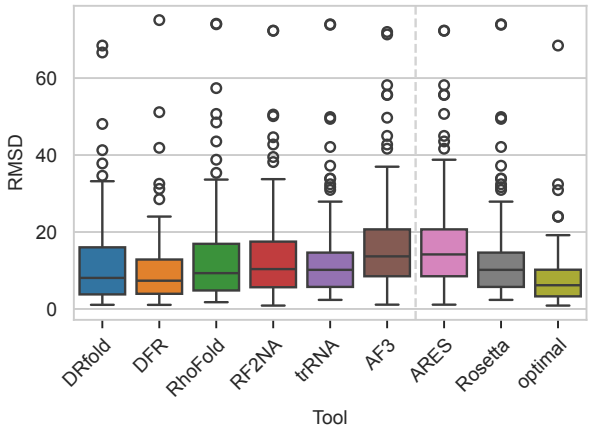

Fig. 1: RMSD for evaluated RNAs across six RNA structure prediction models, ARES, Rosetta score, and an optimal scoring function.

To assess generalization more rigorously, we evaluated model performance on the GenRNA-Struct dataset, a subset of GenRNA, where RNAs were selected to be both sequentially and structurally distinct from training data. Details on dataset construction and additional results are provided in Appendix C. Median RMSD values increased for all models, with DeepFoldRNA rising to 9.97 Å (+3.42%) and AlphaFold 3 to 18.66 Å (+14.27%), confirming that models struggle with structurally novel RNAs and reinforcing their limited generalization beyond training distributions. The results obtained using other structural evaluation metrics further support these findings, with particularly poor performance in non-Watson-Crick interactions (Appendix D).

No single model consistently outperforms others on either GenRNA or GenRNA-Struct datasets, with each producing the best prediction for some RNAs. More specifically, the results on GenRNA dataset showed that DeepFoldRNA leads in 31% of cases, followed by DRfold (22.6%), RhoFold (15.5%), AlphaFold 3 (11.9%), RoseTTAFoldNA (9.5%), and trRosettaRNA (9.5%). This variability suggests that selecting the best prediction on a case-by-case basis could improve overall structure accuracy.

To test this assumption, we investigated whether ARES and Rosetta scoring functions could identify the best predictions for GenRNA dataset. An effective scoring function should rank lower-RMSD structures higher. However, both ARES and Rosetta score performed poorly, showing weak correlation with RMSD (Spearman values: ARES = -0.2190, Rosetta score = -0.0027). ARES strongly favored AlphaFold 3's predictions (83.33% of RNAs), despite AlphaFold 3 having the worst median RMSD on GenRNA. This suggests that ARES may not be effectively evaluating structural accuracy but instead favoring AlphaFold 3 due to training biases. Since ARES was trained exclusively on FARFAR2-generated structures, which differ from deep learning-based predictions, it may struggle to generalize to the structures produced by these models. Retraining ARES on a broader dataset that includes deep learning-based predictions could improve its effectiveness. Similarly, Rosetta score exclusively selected trRosettaRNA's predictions, likely because trRosettaRNA's refinements are guided by Rosetta's own energy function, making its outputs inherently more compatible with Rosetta scoring. This suggests that Rosetta score may prioritize energy-based refinements over structural accuracy, limiting its usefulness as a general-purpose ranking function.

In contrast, an optimal scoring function, one that always selects the prediction with the lowest RMSD, could further reduce the median RMSD from 9.64 Å achieved by the best individual model, DeepFoldRNA, to 7.86 Å. This demonstrates the potential for significant improvement if a more effective scoring approach were developed. Developing more robust scoring methods could improve RNA structure prediction by enabling better model selection beyond individual rankings.

4. Conclusion

This study benchmarks six state-of-the-art deep learning models for RNA structure prediction, highlighting their limited generalization. While each model performs the best for some RNAs, none consistently outperforms the others. Moreover, RMSD values remain above generally acceptable threshold. Performance further drops on structurally novel RNAs, where RMSD values are significantly higher than those observed for RNA Puzzles, showing the models' strong reliance on training-set similarity rather than structural inference.

Beyond the need for better generalization strategies, our results show a weak correlation between ARES and Rosetta scoring functions with RMSD, indicating that current scoring functions do not reliably identify the most accurate RNA 3D structure among model predictions. Finally, we demonstrate that an optimal scoring function could substantially reduce overall RMSD across the evaluation dataset, revealing the potential for significant improvement in existing scoring functions.

Future work should focus on both enhancing model generalization to unseen RNA structures and developing more effective scoring functions that can reliably distinguish high-quality predictions across diverse RNA types, ultimately improving the selection of accurate 3D structures.

Acknowledgments

The authors thank Robin Pearce for assistance with running DeepFoldRNA and for valuable feedback on this work. We also appreciate Rafael Josip Penić for insightful discussions and Iva Bojić for reviewing the manuscript and providing helpful suggestions.

Funding. This work was supported in part by the National Research Foundation (NRF) Competitive Research Programme (CRP) under Project *Identifying Functional RNA Tertiary Structures in Dengue Virus* [NRF-CRP27-2021RS-0001]; in part by Agency for Science, Technology and Research (A*STAR) Industry Alignment Fund - Pre-positioning Programme (IAF-PP) under Project *Drugging RNA: Our Goal for Oncology (DRAGON)* [H23J2a0094]; and in part by the Croatian Science Foundation under Project *Deep Learning-Based RNA Tertiary Structure Prediction and Generation* [IP-2024-05-1554].

References

[1] Run-Wen Yao, Yang Wang, and Ling-Ling Chen. Cellular functions of long noncoding RNAs. *Nature Cell Biology*, 21(5):542–551, May 2019.

[2] Ewen Callaway. Revolutionary cryo-EM is taking over structural biology. *Nature*, 578(7794):201–202, 2020.

[3] Michal J Boniecki, Grzegorz Lach, Wayne K Dawson, Konrad Tomala, Pawel Lukasz, Tomasz Soltysinski, Kristian M Rother, and Janusz M Bujnicki. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic acids research*, 44(7):e63–e63, 2016.

[4] Andrey Krokhotin, Kevin Houlihan, and Nikolay V Dokholyan. iFoldRNA v2: folding RNA with constraints. *Bioinformatics*, 31(17):2891–2893, 2015.

[5] Andrew Martin Watkins, Ramya Rangan, and Rhiju Das. FARFAR2: improved de novo rosetta prediction of complex global RNA folds. *Structure*, 28(8):963–976, 2020.

[6] Marcin Biesiada, Katarzyna Pachulska-Wieczorek, Ryszard W Adamiak, and Katarzyna J Purzycka. RNAComposer and RNA 3D structure prediction for nanotechnology. *Methods*, 103:120–127, 2016.

[7] Yunjie Zhao, Yangyu Huang, Zhou Gong, Yanjie Wang, Jianfen Man, and Yi Xiao. Automated and fast building of three-dimensional RNA structures. *Scientific reports*, 2(1):734, 2012.

[8] Xiaojun Xu, Peinan Zhao, and Shi-Jie Chen. Vfold: a web server for rna structure and folding thermodynamics prediction. *PloS one*, 9(9):e107504, 2014.

[9] Magdalena Rother, Kristian Rother, Tomasz Puton, and Janusz M Bujnicki. ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic acids research*, 39(10):4007–4022, 2011.

[10] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[11] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, Aug 2021.

[12] Yang Li, Chengxin Zhang, Chenjie Feng, Robin Pearce, P. Lydia Freddolino, and Yang Zhang. Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction. *Nature Communications*, 14(1):5745, Sep 2023.

[13] Robin Pearce, Gilbert S Omenn, and Yang Zhang. De novo RNA tertiary structure prediction at atomic resolution using geometric potentials from deep learning. *bioRxiv*, pages 2022–05, 2022.

[14] Congzhou M Sha, Jian Wang, and Nikolay V Dokholyan. Predicting 3D RNA structure from the nucleotide sequence using euclidean neural networks. *Biophys J*, October 2023.

[15] Yuki Kagaya, Zicong Zhang, Nabil Ibtehaz, Xiao Wang, Tsukasa Nakamura, Pranav Deep Punuru, and Daisuke Kihara. NuFold: end-to-end approach for RNA tertiary structure prediction with flexible nucleobase center representation. *Nature communications*, 16(1):881, 2025.

[16] Tao Shen, Zhihang Hu, Zhangzhi Peng, Jiayang Chen, Peng Xiong, Liang Hong, Liangzhen Zheng, Yixuan Wang, Irwin King, Sheng Wang, et al. E2Efold-3D: end-to-end deep learning method for accurate de novo RNA 3D structure prediction. *arXiv preprint arXiv:2207.01586*, 2022.

[17] Minkyung Baek, Ryan McHugh, Ivan Anishchenko, Hanlun Jiang, David Baker, and Frank DiMaio. Accurate prediction of protein-nucleic acid complexes using RoseTTAFoldNA. *Nature Methods*, 21(1):117–121, Jan 2024.

[18] Zongyang Du, Hong Su, Wenkai Wang, Lisha Ye, Hong Wei, Zhenling Peng, Ivan Anishchenko, David Baker, and Jianyi Yang. The trRosetta server for fast and accurate protein structure prediction. *Nature protocols*, 16(12):5634–5651, 2021.

[19] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, pages 1–3, 2024.

[20] Jinsong Zhang, Yuhan Fei, Lei Sun, and Qiangfeng Cliff Zhang. Advances and opportunities in RNA structure experimental determination and computational modeling. *Nature methods*, 19(10):1193–1207, 2022.

[21] Xunxun Wang, Shixiong Yu, En Lou, Ya-Lan Tan, and Zhi-Jie Tan. RNA 3D structure prediction: progress and perspective. *Molecules*, 28(14):5532, 2023.

[22] Jun Zhang, Mei Lang, Yaoqi Zhou, and Yang Zhang. Predicting RNA structures and functions by artificial intelligence. *Trends in Genetics*, 2024.

[23] Sunandan Mukherjee, S Naeim Moafinejad, Nagendar Goud Badepally, Katarzyna Merdas, and Janusz M Bujnicki. Advances in the field of RNA 3D structure prediction and modeling, with purely theoretical approaches, and with the use of experimental data. *Structure*, 2024.

[24] Akash Bahai, Chee Keong Kwoh, Yuguang Mu, and Yinghui Li. Systematic benchmarking of deep-learning methods for tertiary RNA structure prediction. *bioRxiv*, pages 2024–02, 2024.

[25] Clément Bernard, Guillaume Postic, Sahar Ghannay, and Fariza Tahi. State-of-the-RNArt: benchmarking current methods for RNA 3D structure prediction. *NAR Genomics and Bioinformatics*, 6(2):lqae048, 2024.

[26] Clément Bernard, Guillaume Postic, Sahar Ghannay, and Fariza Tahi. Has AlphaFold 3 reached its success for RNAs? *bioRxiv*, pages 2024–06, 2024.

[27] Chandran Nithin, Sebastian Kmiecik, Roman Błaszczyk, Julita Nowicka, and Irina Tuszyńska. Comparative analysis of RNA 3D structure prediction methods: towards enhanced modeling of RNA–ligand interactions. *Nucleic Acids Research*, 52(13):7465–7486, 2024.

[28] Sumit Tarafder, Rahmatullah Roche, and Debswapna Bhattacharya. The landscape of RNA 3D structure modeling with transformer networks. *Biology Methods and Protocols*, 9(1):bpae047, 2024.

[29] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[30] Jun Li, Wei Zhu, Jun Wang, Wenfei Li, Sheng Gong, Jian Zhang, and Wei Wang. RNA3DCNN: Local and global quality assessments of RNA 3D structures using 3D deep convolutional neural networks. *PLoS computational biology*, 14(11):e1006514, 2018.

[31] Sumit Tarafder and Debswapna Bhattacharya. lociPARSE: a locality-aware invariant point attention model for scoring RNA 3D structures. *bioRxiv*, 2023.

[32] Ya-Lan Tan, Xunxun Wang, Shixiong Yu, Bengong Zhang, and Zhi-Jie Tan. cgRNASP: coarse-grained statistical potentials with residue separation for RNA structure evaluation. *NAR Genomics and Bioinformatics*, 5(1):lqad016, 2023.

[33] Raphael JL Townshend, Stephan Eismann, Andrew M Watkins, Ramya Rangan, Masha Karelina, Rhiju Das, and Ron O Dror. Geometric deep learning of RNA structure. *Science*, 373(6558):1047–1051, 2021.

[34] Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J O'Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031–3048, 2017.

[35] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, January 2000.

[36] Marcin Magnus, Maciej Antczak, Tomasz Zok, Jakub Wiedemann, Piotr Lukasiak, Yang Cao, Janusz M Bujnicki, Eric Westhof, Marta Szachniuk, and Zhichao Miao. RNA-Puzzles toolkit: a computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools. *Nucleic Acids Res*, 48(2):576–588, January 2020.

[37] José A. Cruz, Marc-Frédérick Blanchet, Michal Boniecki, Janusz M. Bujnicki, Shi-Jie Chen, Song Cao, Rhiju Das, Feng Ding, Nikolay V. Dokholyan, Samuel C. Flores, et al. RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA (Cambridge)*, 18(4):610–625, 2012.

[38] Zhichao Miao, Ryszard W. Adamiak, Marc-Frédérick Blanchet, Michal Boniecki, Janusz M. Bujnicki, Shi-Jie Chen, Clarence Cheng, Grzegorz Chojnowski, Fang-Chieh Chou, Pablo Cordero, et al. RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA (Cambridge)*, 21(6):1066–1084, 2015.

[39] Zhichao Miao, Ryszard W. Adamiak, Maciej Antczak, Robert T. Batey, Alexander J. Becka, Marcin Biesiada, Michał J. Boniecki, Janusz M. Bujnicki, Shi-Jie Chen, Clarence Y. Cheng, et al. RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA (Cambridge)*, 23(5):655–672, 2017.

[40] Zhichao Miao, Ryszard W Adamiak, Maciej Antczak, Michał J Boniecki, Janusz Bujnicki, Shi-Jie Chen, Clarence Yu Cheng, Yi Cheng, Fang-Chieh Chou, Rhiju Das, et al. RNA-Puzzles Round IV: 3D structure predictions of four ribozymes and two aptamers. *RNA*, 26(8):982–995, 2020.

[41] Fan Bu, Yagoub Adam, Ryszard W Adamiak, Maciej Antczak, Belisa Rebeca H de Aquino, Nagendar Goud Badepally, Robert T Batey, Eugene F Baulin, Pawel Boinski, Michal J Boniecki, et al. RNA-Puzzles Round V: blind predictions of 23 RNA structures. *Nature methods*, 22(2):399–411, 2025.

[42] Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, November 2017.

[43] Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nature Communications*, 9(1):2542, June 2018.

[44] Rachael C Kretsch, Ebbe S Andersen, Janusz M Bujnicki, Wah Chiu, Rhiju Das, Bingnan Luo, Benoît Masquida, Ewan KS McRae, Griffin M Schroeder, Zhaoming Su, et al. RNA target highlights in CASP15: Evaluation of predicted models by structure providers. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1600–1615, 2023.

[45] Marcell Szikszai, Marcin Magnus, Siddhant Sanghi, Sachin Kadyan, Nazim Bouatta, and Elena Rivas. RNA3DB: A structurally-dissimilar dataset split for training and benchmarking deep learning models for RNA structure prediction. *Journal of Molecular Biology*, page 168552, 2024.

[46] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

[47] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.

[48] Chengxin Zhang, Morgan Shine, Anna Marie Pyle, and Yang Zhang. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature Methods*, 19(9):1109–1115, Sep 2022.

[49] Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

[50] Marco Biasini, Valerio Mariani, Jürgen Haas, Stefan Scheuber, Andreas D Schenk, Torsten Schwede, and Ansgar Philippsen. OpenStructure: a flexible software framework for computational structural biology. *Bioinformatics*, 26(20):2626–2628, 2010.

[51] Vincent B Chen, W Bryan Arendall, Jeffrey J Headd, Daniel A Keedy, Robert M Immormino, Gary J Kapral, Laura W Murray, Jane S Richardson, and David C Richardson. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 66(1):12–21, 2010.

[52] Marc Parisien, José Almeida Cruz, Éric Westhof, and François Major. New metrics for comparing and assessing discrepancies between rna 3d structures and models. *Rna*, 15(10):1875–1885, 2009.

[53] Patrick Gendron, Sébastien Lemieux, and François Major. Quantitative analysis of nucleic acid three-dimensional structures. *Journal of molecular biology*, 308(5):919–936, 2001.

Appendix A. GenRNA dataset creation

Our aim was to create GenRNA, a dataset that could effectively evaluate the generalization abilities of the models. Since most models do not explicitly disclose their training datasets, we used 13 January 2023, the validation set cutoff date for AlphaFold 3, as the training cutoff for all models. This assumption is reasonable, as five other models were developed for CASP15 in April 2022, indicating they were trained on data published well before our chosen cutoff date. To compile this dataset, we selected structures published in the PDB database after 13 January 2023, ensuring that these RNAs were unseen during training. This approach enables a robust assessment of the models' generalization ability to novel structures.

Starting with a download of all available RNA chains from PDB, which was $20,320$ RNA chains, we applied sequence identity clustering using MMseqs2 [42, 43], filtering for a minimum sequence identity of 90% and coverage of at least 80%, resulting in $3,822$ clusters. Clusters containing only RNAs published after 13 January 2023 were retained, followed by further filtering to remove sequences shorter than 16 nucleotides, with resolution greater than 9 Å, those with fewer than 90% defined residues, and those consisting solely of unknown residues ('N' or 'X'). After these steps, 143 clusters remained, each uniquely represented by sequences with optimal resolution

and maximal percentage of nucleotides with defined locations. This process is visualized in Figure A1.

For comparability, we only included sequences for which all models could produce final structures, yielding a final subset of 84 sequences.

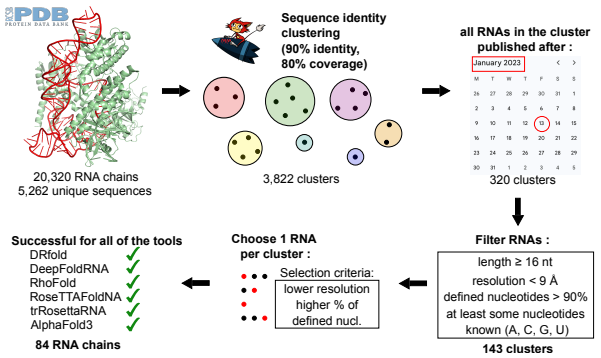

Fig. A1: Process of creating the evaluation dataset.

Appendix B. RNA Puzzles Dataset

The RNA Puzzles dataset consists of 37 RNA structures from the RNA-Puzzles initiative [37–41], a widely used benchmark for RNA 3D structure prediction. To avoid overlap with CASP15 targets [44], Puzzles 35 and 36 were excluded. While it is unknown whether these RNA Puzzles were explicitly included in the training datasets of the evaluated models, many were published in the Protein Data Bank (PDB) well before the models were developed. As a result, models may have encountered structurally similar RNAs during training, potentially making this dataset less challenging for evaluation.

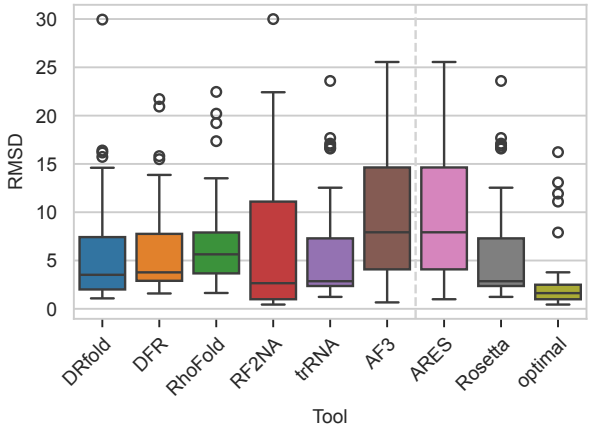

Fig. A2: RMSD for RNA Puzzles across six RNA structure prediction models, ARES, Rosetta score, and an optimal scoring function.

When evaluated on this dataset, models achieve their best performance, with significantly lower RMSD values compared to our dataset. RoseTTAFoldNA achieves the lowest median RMSD (2.65 Å), indicating that these models perform particularly well when trained on similar examples. AlphaFold 3, however, lags behind, with the highest median RMSD (7.92Å), suggesting differences in training data or model architecture. These results reinforce that models struggle more when applied to unseen RNAs, as performance drops significantly when evaluated on datasets without training-set similarity.

Appendix C. GenRNA-Struct: Stricter generalization assessment using structural and sequential clustering

To further assess generalization, we constructed GenRNA-Struct, a subset of GenRNA, designed to include RNAs that are not only sequentially distinct from training data but also structurally dissimilar. We applied the RNA3DB pipeline [45], which clusters RNAs based on both sequence and structural similarity. We retained only RNAs from clusters where all members were published after 13 January 2023, ensuring that no structural relatives were present in training data. The final evaluation subset contains 60 RNAs, forming a stricter benchmark for model generalization.

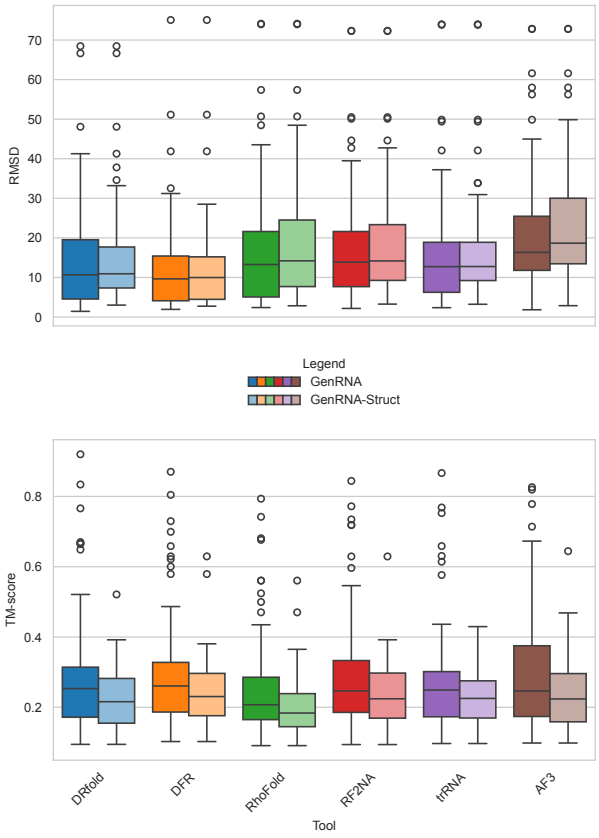

Fig. A3: RMSD (upper panel) and TM-score (lower panel) for GenRNA and GenRNA-Struct datasets.

Figure A3 shows comparison of all-atom RMSD and TM-score [46] performance on GenRNA and

GenRNA-Struct datasets. We observe that median RMSD remains similar, but TM-scores are slightly lower, with all models continuing to produce very low TM-scores (~ 0.2), consistent with random folds.

An example of these modeling challenges is shown in Figure A4, which illustrates predictions for the RNA chain with PDB_ID 8T29_R, included in both GenRNA and GenRNA-Struct. The native structure is shown in green and the predicted structures in blue. All six predicted structures for this RNA chain exhibit relatively high RMSD and low TM-scores. Among these, AlphaFold 3 produced the most accurate prediction, capturing a reasonable global fold. Visualizations of the structures were generated using PyMOL [47].

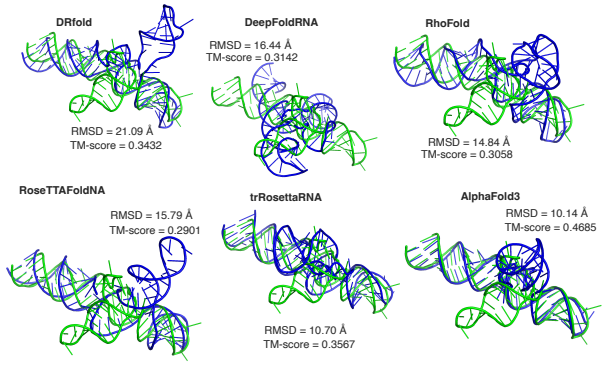

Fig. A4: Predictions for RNA chain 8T29_R.

Appendix D. Other metrics

In addition to all-atom RMSD, we calculate seven additional metrics to evaluate RNA 3D structure predictions. TM-score [46], computed using USalign [48], assesses how well the predicted global fold aligns with the native structure. lDDT [49], calculated using OpenStructure [50], evaluates local atomic accuracy. The clash score [51], computed with MolProbity [51], measures steric clashes between atoms, where lower values indicate better structural feasibility. Interaction Network Fidelity (INF) metrics [52], calculated using the RNA Puzzles Toolkit [36], assess RNA-specific structural accuracy by detecting essential base interactions with MC-Annotate [53]. These include INF_WC (Watson-Crick interactions), INF_NWC (non-Watson-Crick interactions), and INF_STACK (stacking interactions), as well as an overall INF_ALL score.

Figure A5 shows the median values for each tool across the GenRNA dataset. Since all metrics except RMSD and clash score range from 0 to 1, where 1 represents perfect agreement with the native structure, we adjusted RMSD and clash score for consistency. Specifically, we took their negative values (as lower values are better for these metrics) and applied min-max normalization, ensuring that a larger surface area in the spider plot indicates better overall performance.

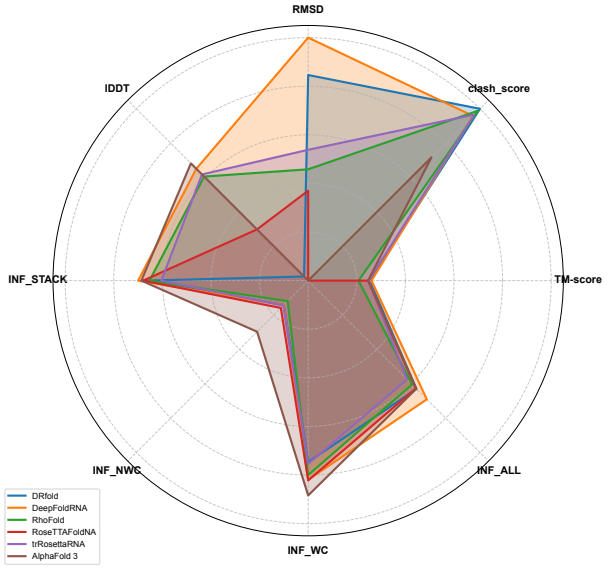

Fig. A5: Comparison of tool performance across normalized metrics on GenRNA.

TM-score results indicate that none of the models achieve meaningful global fold accuracy, with median values ranging from 0.207 for RhoFold to 0.261 for DeepFoldRNA, which corresponds to randomly generated folds since only values above 0.45 indicate meaningful structural similarity. While INF_WC and INF_STACK show moderate agreement with native interactions, INF_NWC remains particularly challenging, with all models performing poorly. Among them, AlphaFold 3 achieves the highest INF_NWC score, but the overall accuracy of non-Watson-Crick interaction predictions remains low across all models.

No single tool performs best across all metrics: DeepFoldRNA achieves the best results for all-atom RMSD, TM-score, INF_ALL, and INF_STACK, while also maintaining a respectable clash score (third best, median = 0.76) and strong lDDT (second best, median = 0.651). However, it performs worse for INF_WC and INF_NWC, where AlphaFold 3 and RoseTTAFoldNA outperform it.

Figure A6 presents the same evaluation metrics for GenRNA-Struct, a stricter subset of GenRNA. Given that GenRNA-Struct comprises over 70% of GenRNA, the overall trends remain consistent. DeepFoldRNA continues to achieve the best performance across most metrics. However, the results still fall short of desirable accuracy, with TM-scores only slightly above 0.2. While some INF metrics, such as INF_WC, show strong agreement with native interactions (median > 0.84), INF_NWC remains a major challenge across all models.

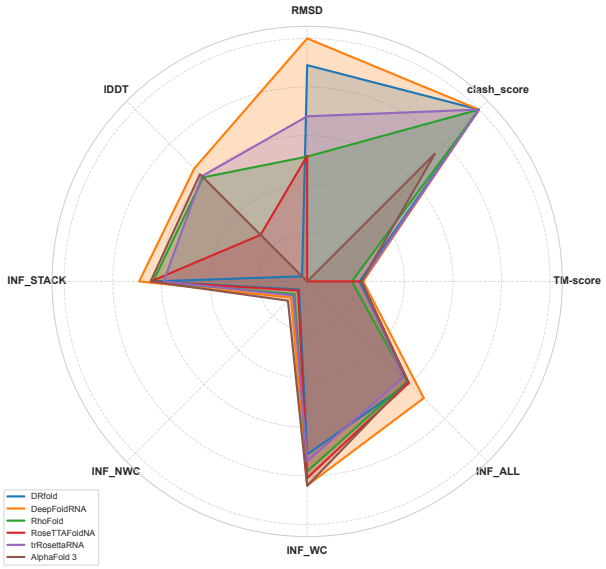

Fig. A6: Comparison of tool performance across normalized metrics on GenRNA-Struct.