Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

From sequence to structure: A comprehensive review of deep learning models for RNA structure prediction

Utkarsh Upadhyay, Anton Dorn, Christian Faber and Alexander Schug[1]

RNA structure prediction remains one of the most challenging problems in computational biology, with significant implications for understanding gene regulation, drug design, and synthetic biology. While deep learning has revolutionized protein structure prediction, RNA presents unique challenges including limited training data, complex noncanonical interactions, and conformational flexibility. This review examines the evolution from traditional physics-based methods to current deep learning approaches for RNA secondary and tertiary structure prediction. After briefly exploring traditional methods, like Direct Coupling Analysis and physics-based simulations, we systematically review three deep learning paradigms: language model–based methods, end-to-end structure predictors, and geometry-distance prediction approaches. Furthermore, we identify critical future research directions focusing on advanced tokenization strategies to address data scarcity and explainable artificial intelligence techniques to improve model interpretability. Despite significant progress, achieving transformative performance requires continued methodological innovation, specifically designed for RNA's unique characteristics, and a substantial expansion of high-quality structural datasets.

Addresses
Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany

Corresponding author: Schug, Alexander. (alexander.schug@kit.edu)
[1] Current address: Scientific Computing Center, Karlsruhe Institute of Technology, Germany

Current Opinion in Structural Biology 2026, **97**:103216

This review comes from a themed issue on **Artificial Intelligence (AI) Methodologies in Structural Biology (2026)**

Edited by **Yana Shen** and **Jianyi Yang**

For complete overview of the section, please refer the article collection - Artificial Intelligence (AI) Methodologies in Structural Biology (2026)

Available online 5 February 2026

https://doi.org/10.1016/j.sbi.2025.103216

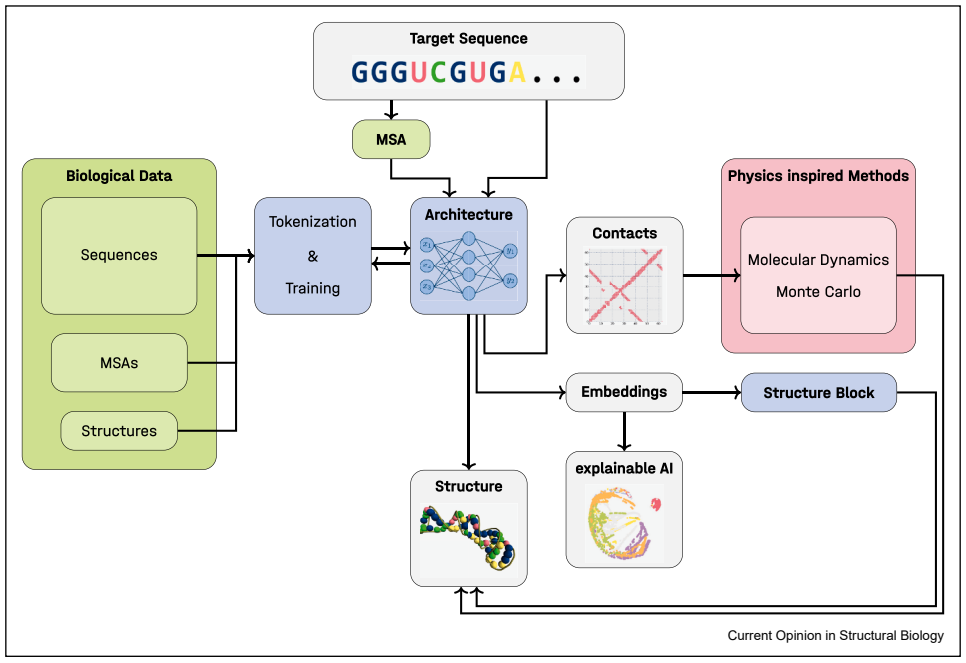
Introduction

RNA molecules play crucial roles in a wide range of biological processes, including transcription regulation, cell signaling, catalysis, and posttranscriptional control, with their diverse functions deeply related to their structures [1–3]. Experimental methods like nuclear magnetic resonance, X-ray crystallography, and cryogenic electron microscopy (cryo-EM) can provide detailed structural information [4]. The experimental study of RNA is a large and very active field of research as evidenced by many significant findings [5,6]. However, they are resource-intensive and face size and resolution limits due to RNA's high flexibility, conformational heterogeneity, and tendency to resist crystallization. Unlike proteins, where remarkable computational advances have been achieved, RNA structure prediction continues to present unique challenges (See Figure 1).

Computational RNA structure prediction has evolved through distinct methodological phases. Traditional approaches employed physics-inspired, coarse-grained models using Monte Carlo sampling, exemplified by SimRNA [7], which leverages a statistical potential to explore conformational space, and *de novo* fragment assembly methods like FARFAR2, which integrates fragment libraries and helix modeling to achieve accurate predictions of native-like RNA tertiary folds [8]. A critical breakthrough came with Direct Coupling Analysis (DCA) [9,10], a statistical inference method that identifies co-evolving nucleotide pairs from multiple sequence alignments (MSAs). DCA applies inverse Potts models and maximum entropy principles to infer direct evolutionary couplings, which can then serve as distance constraints in tertiary structure modeling.

The advent of deep learning revolutionized the field, motivated by success in protein structure prediction. Early deep learning approaches for RNA included methods like RNAContact [11] and CoCoNet [12], which employed convolutional neural networks and feature engineering strategies to work effectively with limited annotated structural data.

Figure 1



Workflow diagram showing the diverse computational pathways for RNA structure prediction. Green components represent biological data sources with varying availability: sequences (most abundant), multiple sequence alignments (moderate availability), and experimental structures (most scarce). **Blue components** show the core deep learning pipeline: (1) Tokenization converts raw sequences into representations, while training involves self-supervised pretraining on large sequence datasets (language models), supervised training on structural data (end-to-end predictors), or multitask training combining sequence and structure objectives. (2) Architecture encompasses the neural network designs (transformers, CNNs, and diffusion models). (3) Contacts represent intermediate structural predictions (distance maps and contact probabilities) generated by geometry-based approaches. (4) Embeddings are high-dimensional representations from language models that capture structural patterns. (5) Structure block performs the final 3D coordinate prediction or reconstruction. **Red components** indicate physics-inspired methods that can augment deep learning predictions by providing them with energy-based refinement. Explainable AI (bottom right) represents emerging efforts to interpret model decisions through attention visualization, saliency analysis, and feature attribution. The diagram emphasizes that modern approaches typically combine multiple pathways. The workflow accommodates both sequence-only methods and MSA-dependent approaches, highlighting the field's methodological diversity in addressing RNA's unique structural prediction challenges. AI, artificial intelligence; CNNs, convolutional neural networks; MSA, multiple sequence alignment.

Building on these foundations, recent years have seen a shift toward self-supervised techniques based on transformer architectures. For example, BARNACLE [13] uses pretraining on MSAs of RNA families to achieve excellent performance on the contact map prediction task. Most recently, the focus has shifted toward large language models trained directly on sequence data, which can learn structural representations without requiring explicit evolutionary information.

Despite these advances, RNA structure prediction faces distinct challenges that have prevented achievements comparable to those seen for proteins. Models for protein structure prediction [14] achieve sub-2 Å accuracy for most protein targets as they have more than 200,000 available protein structures for training, while RNA methods work with around 9000 RNA structures representing less than 1 % of Protein Data Bank (PDB) depositions and originating from only about 100 distinct RNA families. These fundamental limitations and their impact on RNA structure prediction have been comprehensively analyzed in recent reviews [15,16].

This review provides a comprehensive analysis of the most recent advances in deep learning methods for RNA structure prediction. We discuss the datasets driving these artificial intelligence (AI) models, summarize cutting-edge computational methods, analyze their performance and limitations, and identify promising directions for methodological advancement. Our goal is to provide computational biologists with essential insights into current capabilities and future opportunities in this rapidly evolving field, while acknowledging the persistent challenges that distinguish RNA structure prediction from other successful applications of deep learning in structural biology.

Challenges in RNA structure prediction

RNA structure prediction faces fundamentally different computational challenges compared to protein structure

prediction, creating intractable problems for traditional algorithms and limiting the effectiveness of deep learning approaches. These challenges stem from RNA's unique structural characteristics and complex folding patterns [17] that distinguish it from protein folding mechanisms.

RNA exhibits significantly greater structural flexibility because its backbone is defined by eight torsional angles compared to proteins' simpler backbone geometry (three torsion angles), creating a vastly larger conformational landscape [18]. Many functional RNAs adopt multiple conformational states—riboswitches switch between conformations depending on ligand binding, viral RNAs must refold at different life cycle stages, and regulatory RNAs switch between active and inactive conformations [19,20]. This conformational flexibility presents a fundamental challenge for single-structure prediction approaches as the folding landscape may be better represented as multiple low-energy wells rather than a single energy minimum.

In addition to the conformational challenges, RNA exhibits diverse base pairing interactions. Beyond Watson–Crick base pairs, noncanonical base pairs constitute approximately 33 % of all base pairs in structured RNAs. The Leontis–Westhof [21] classification defines 12 distinct base pair families comprising 192 base pair classes when considering geometric types and sequence context, including Hoogsteen pairs, sugar–edge interactions, and other non-standard hydrogen bonding patterns [22,23].

In addition, there are topological challenges in base pairing, such as pseudoknots, occurring when bases in different loops pair with each other to form non-nested structures that violate standard nesting conventions.

However, by far the greatest challenge for prediction using AI methods is the scarcity of experimental data (Table 1). As of September 2025, RNA-only structures comprise less than 1 % of the approximately 230,000 structures in the PDB, with a strong bias toward simpler structures like tRNAs and rRNA subunits. This leads to insufficient structural diversity for training robust machine learning models. Tools like RNAsolo [24], RNA3DB [25], and NucleoSeeker [26] provide well-curated datasets for RNA but isolated RNA structures are still scarce, thus limiting the training data.

Traditional methods for RNA structure prediction

Traditional RNA structure prediction relies on computationally intensive simulation software such as SimRNA [7], FARFAR2 [8], or molecular dynamics simulations. As these methods often yield insufficient accuracy when used alone, additional constraints are commonly incorporated into the prediction process. Contact predictions, typically derived from evolutionary analysis, are frequently added as distance constraints to guide structure prediction.

DCA[9,10] represents a well-known approach for contact prediction, employing maximum likelihood estimation to infer conserved residues and predict contact maps from MSAs. These self-supervised techniques leverage evolutionary information encoded in MSAs without requiring labeled datasets, making them particularly valuable when experimental structural data is limited. However, their effectiveness is constrained by the availability of high-quality MSAs, which are often scarce for RNA families due to limited sequence diversity and poor alignment quality.

Physics-based methods employ molecular dynamics simulations and Monte Carlo sampling to explore RNA conformational space using empirical force fields derived from experimental data and quantum mechanical calculations. Fragment assembly techniques like FARFAR2 construct structures by combining short RNA fragments from experimental structures through iterative assembly and energy minimization. While these approaches provide detailed atomic models with explicit consideration of physical interactions and remain competitive in blind prediction challenges [16], they face significant computational limitations. Adequate sampling of the conformational space for large RNAs requires substantial computational resources, limiting their applicability to high-throughput or genome-scale predictions. Deep learning methods aim to achieve comparable accuracy with orders-of-magnitude reduction in computational time.

Hybrid and experimental integration approaches

Combined approaches integrate multiple methodological strategies to overcome the limitations of individual techniques, representing an emerging and diverse landscape of RNA structure prediction methods. These approaches range from traditional combinations of computational methods [35] to AI-enhanced experimental techniques [36] and physics-based innovations [37,38].

An emerging field involves the development of machine-learned force fields (MLFFs) for molecular dynamics simulations, where traditional parameters manually fitted to experimental data or quantum mechanical calculations are replaced with AI-optimized versions. While these efforts have concentrated primarily on protein structure prediction, some generalist approaches include RNA in their frameworks. Recent examples include espaloma-0.3 [39] and Grappa [40], both deploying graph neural networks trained on

Table 1

Comprehensive overview of datasets and benchmarks driving RNA structure prediction research across primary sequence (1D), secondary structure (2D), and tertiary structure (3D) levels. Current size indicates the number of structures or sequences as of the last update. Update frequency reflects the maintenance commitment of each resource, ranging from static benchmarks to frequently updated repositories. Source methodology distinguishes between comparative analysis using homology and covariance models and other curation methods. Key features highlight the distinguishing characteristics and specialized capabilities of each dataset, including structural diversity measures, family coverage, and specific focus areas such as challenging structural motifs or blind prediction targets. NucleoSeeker entry differs from other resources as it represents a dynamic curation tool: 'PDB-derived' indicates variable dataset sizes depending on user-specified filters, 'On-demand' reflects that it always uses the latest PDB data.

Dataset	Current size	Last update	Update frequency	Source methodology	Key features
Tertiary structure datasets					
RNA3DB [25]	23,185 PDB chains	December 2024	Quarterly	PDB + Rfam filtering, structural clustering	135 structurally dissimilar components
RNAsolo2 [24]	31,928 PDB chains	September 2025	Weekly	PDB cleaning + BGSU classification	4491 representative classes
BGSU representative set [27]	19,528 PDB chains	September 2025	Weekly	IFE-optimized representatives	4593 representative classes
RNA-Puzzles [28]	41 RNA targets	December 2024	Every 2–3 years	Blind prediction targets	Focus on challenging aspects like coaxial stacking and non-Watson–Crick modules
CASP15-RNA [29]	13 RNA targets	December 2022	Biennial competition	Novel structures without homologs	First RNA inclusion in CASP
CASP16-RNA [29]	72 RNA targets	December 2024	Biennial competition	Novel structures including complexes	Significant increase from CASP15
Secondary structure datasets					
bpRNA-1m [30]	102,318 structures	May 2018	Static	7-Database aggregation	Comprehensive feature annotation
RNAStrAlign [31]	30,452 samples	November 2017	Static	Homologous family alignment	Diverse RNA families
ArchiveII [32]	3975 structures	September 2016	Static benchmark	Combination of multiple benchmarks	Entirely different families combined together
Single sequence datasets					
RNAcentral [33]	>40 million sequences	June 2025	Every 3 months	54 Expert databases	All noncoding RNAs
Rfam [34]	4178 families	September 2024	1–2 years	Covariance model families	Most up-to-date collection of RNA families
Dataset curation tools					
NucleoSeeker [26]	PDB-derived	2025	On-demand	Multiple structure and sequence related filters	Customizable redundancy removal, quality-based curation

PDB, Protein Data Bank.

quantum chemical datasets. The development of RNA-specific MLFFs represents a promising future direction that could significantly improve the accuracy of physics-based RNA structure prediction.

Complementing sequence-based approaches, recent developments focus on reconstructing RNA structure from experimental cryo-EM maps rather than the sequence alone. Notable advances include DeepTracer-2.0 [41], CryoREAD [42], and DeepCryoRNA [43], which report significant improvements over existing classical and machine learning approaches.

State-of-the-art deep learning methods

Deep learning approaches for RNA structure prediction can be categorized into three main paradigms: language model–based methods, end-to-end structure predictors, and geometry-distance prediction models, each addressing different aspects of the prediction challenge (Table 2).

Table 2

Technical specifications of various methods for RNA structure prediction, detailing the complete computational pipeline from input processing to final structure generation. AI input specifies the data types each method processes, ranging from single sequences to complex combinations including multiple sequence alignments (MSAs) and secondary structures (SSs). AI output categorizes the direct neural network predictions: embeddings from language models, probability distributions over contacts or folding configurations, and geometric parameters. Model architecture describes the core neural network designs with parameter counts indicating computational scale, encompassing Bidirectional encoder representations from transformers language models (BERT variants), convolutional networks for contact prediction, and specialized architectures like invariant point attention (IPA) for geometric reasoning. Post-processing methods convert neural network outputs to final structures through diverse approaches: simple thresholding, dynamic programming optimization, physics-based energy minimization, and geometric reconstruction algorithms.

Method	AI input	AI output	Model architecture	Postprocessing
		Language models		
RiNALMo [46]	Sequence	Embeddings	BERT-650M	2D ResNet
RNA-FM [44]	Sequence	Embeddings	BERT-99.5M	2D ResNet
NucleicBERT [47]	Sequence	Embeddings	BERT-404M	2D ResNet
RNAErnie [45]	Sequence	Folding scores	BERT-86.7M	Dynamic programming
		End-to-end 3D predictors		
RhoFold+ [48]	Seq + MSA	Frame + torsion	IPA-transformer	Reconstruction
RoseTTAFoldNA [49]	Seq + MSA	Frame + torsion	SE(3)-transformer	Reconstruction
AlphaFold3 [51]	Seq + MSA	Full-atom coordinates	Diffusion	None
DRFold [50]	Seq	Coarse-grained representation	Transformer + IPA	L-BFGS and MD refinement
		Geometry & distance predictors		
trRosettaRNA [52]	Seq + MSA + SS	Distance + geometry	Transformer	Rosetta energy min
BARNACLE [13]	MSA	Contact probability	Transformer + XGBoost	Threshold
UFold [53]	Sequence	Contact probability	U-Net CNN	Constrained optimization
SPOT-RNA2 [55]	Seq + MSA	Contact probability	Dilated CNN	Threshold
RNADiffFold [54]	Sequence	Contact probability	Diffusion + Transformer	Threshold
		Physics-based methods		
FARFAR2 [8]	Seq + constraints	Full-atom coordinates	Fragment assembly	Energy minimization
SimRNA [7]	Seq + constraints	Coarse coordinates	MC sampling	All-atom rebuild

Language model approaches leverage transformer architectures pretrained on large sequence datasets to capture structural patterns. RNA-FM [44], RNAErnie [45], RiNALMo [46], and NucleicBERT [47] represent large language models that learn sequence—structure relationships through self-supervised training on millions of RNA sequences. RNA-FM establishes early benchmarks for RNA language modeling and achieves strong performance. RNAErnie introduces sophisticated motif-aware pretraining with multi-level masking strategies, incorporating biological priors for enhanced structural understanding. RiNALMo is the largest language model trained on RNA sequences. NucleicBERT extends language modeling to diverse RNA tasks while incorporating explainable AI (xAI) analysis for understanding model predictions, addressing the critical interpretability challenges in biological applications. These models generate embeddings that can be processed by downstream networks for various structural and functional tasks.

End-to-end structure prediction methods directly output 3D coordinates or structural representations. RhoFold+ [48] combines a language model with AlphaFold2-inspired architectures, achieving excellent performance on RNA-Puzzles benchmarks. RoseTTAFoldNA [49] extends protein-folding architectures to RNA-protein complexes using SE(3)-equivariant transformers that simultaneously update 1D, 2D, and 3D representations. DRFold [50] does end-to-end coarse-grained prediction with molecular dynamics refinement for post processing. AlphaFold3 [51] represents a paradigm shift by employing diffusion models to directly predict full-atom coordinates for all biomolecules, eliminating the need for backbone frame prediction and side-chain reconstruction procedures used by other methods.

Geometry and distance prediction approaches predict intermediate structural features that are then converted to final structures through optimization algorithms.

BARNACLE [13] employs transformer architectures combined with XGBoost to predict RNA contact probabilities from MSAs. trRosettaRNA [52] predicts distance maps and geometries using transformers and then employs Rosetta energy minimization to generate final structures, achieving superior stereochemistry through physics-based refinement. UFold [53] uses U-Net architectures with constrained optimization for contact probability prediction.

Recent innovations include diffusion-based approaches that model the data distribution of RNA conformations. RNADiffFold [54] applies diffusion models to secondary structure prediction, while AlphaFold3's success suggests broader potential for generative modeling in RNA structure prediction. These methods can potentially capture conformational flexibility and generate multiple structural states, addressing RNA's inherent dynamic nature.

Performance varies significantly across method types and target difficulty. End-to-end methods generally achieve better accuracy on well-folded targets, while geometry-based approaches often provide better stereochemical quality. Language model integration consistently improves performance across all architectures, particularly for sequences with limited evolutionary information, demonstrating the value of self-supervised pre-training in addressing data scarcity challenges.

Despite these advances, several critical considerations affect the interpretation of reported performance metrics across deep learning approaches. Many state-of-the-art methods, including RhoFold+, RoseTTAFoldNA, and trRosettaRNA, rely heavily on evolutionary information, limiting their applicability to sequences lacking sufficient homologs or high-quality alignments. Additionally, rigorous dataset curation using tools like NucleoSeeker [26] has revealed that data leakage between training and testing sets represents an unnoticed challenge in the field. This overlap can lead to overly optimistic performance estimates as models may exploit memorized patterns rather than learning generalizable structural principles. Future method development and evaluation must jointly apply temporal dataset separation and comprehensive redundancy filtering as neither alone is sufficient to ensure robust assessment of true predictive capabilities. Only through such rigorous benchmarking can the field accurately gauge progress toward solving RNA's unique structural prediction challenges.

Future directions

Modern methods demonstrate the potential for solving complex RNA structure prediction problems yet they face significant bottlenecks requiring further optimization.

First, deep learning models are prone to overfitting, where models may learn specific features (including noise and bias) in the training data while ignoring more general patterns. This overfitting occurs primarily due to scarce or noisy data, emphasizing the need for large-scale, high-quality datasets. Advanced tokenization strategies help address RNA's data scarcity as biologically informed token representations can enable models to extract more structural information from sequences compared to single-nucleotide approaches. While most current models employ a single-nucleotide tokenization (RNA-FM [44] and RiNALMo [46]), specialized approaches have emerged: adaptive dual tokenization combining nucleotide and byte-pair encoding (BiRNA-BERT [56]), and convolutional encoding for long sequences (lncRNA-BERT [57]). RNAErnie [45] introduces sophisticated motif-aware pretraining with three-level masking: base-level masking (15 % nucleotides), subsequence-level masking (4−8 bp segments), and motif-level masking using biological priors. Multi-species training approaches show increasing sophistication, with SpliceBERT [58] training on 2M + precursor mRNA sequences from 72 vertebrates for evolutionary conservation detection and cross-species splice site prediction. These tokenization advances highlight the critical importance of RNA-specific linguistic representations in achieving improved structural prediction accuracy.

Second, the "black-box" nature of deep learning models leads to a critical lack of interpretability, particularly problematic given the high complexity of RNA structures and data scarcity, which increases prediction uncertainty. This interpretability gap is especially critical in biological applications, where researchers need to understand the biological mechanisms behind predictions for experimental validation and therapeutic development. Advanced xAI techniques offer promising solutions, including saliency analysis using GradCAM [59] for visualizing important sequence regions, analyzing attention weights in transformer layers to understand long-range dependencies [60], integrated gradients for attributing predictions to specific nucleotides, and SHAP (SHapley Additive exPlanations) [61] values for quantifying feature contributions to structural predictions.

Future breakthroughs require addressing data scarcity through 10−50x expansion of high-quality RNA structures across diverse families, addressing current taxonomic and functional biases while including dynamic structural information, and context-dependent structures under different ionic conditions and binding states. Integration of experimental constraints such as chemical probing techniques (selective 2'-hydroxyl acylation by primer expansion (SHAPE), dimethyl sulfate (DMS) based techniques, and cross-linking) and cryo-EM studies represents a promising

avenue for improving prediction accuracy through hybrid computational-experimental approaches. The development of ensemble-based methods capable of predicting multiple conformational states remains essential for capturing RNA's inherent flexibility and functional diversity.

RNA structure prediction represents a multidisciplinary field spanning computational biology, structural biology, and AI, requiring coordinated collaborative efforts and open-source initiatives to drive rapid development. Expert consensus suggests that achieving RNA's "AlphaFold moment" will require 5–10 years of coordinated efforts in experimental structure determination, algorithmic innovation specifically designed for RNA's unique characteristics, and community benchmarking infrastructure development [62]. Success depends on recognizing that RNA structure prediction faces qualitatively different challenges from protein folding, demanding RNA-specific solutions rather than direct adaptation of protein methods.

Data availability

No new data were created or analyzed during this study. Data sharing is not applicable to this article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

A.S. recognizes support by HIDSS4Health—the Helmholtz Information & Data Science School for Health. A.S. and A.D. recognize support by the Helmholtz Association's Initiative and Networking Fund (INF) under the grant MadRNA. A.S. and C.F. recognize support by the Helmholtz Foundation Model Initiative (HFMI) of the Helmholtz Association under the grants PROFOUND and Virtual Cell. The funders had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. Mattick JS, *et al.*: **Long non-coding RNAs: definitions, functions, challenges and recommendations**. *Nat Rev Mol Cell Biol* 2023 Jun, **24**:430–447, https://doi.org/10.1038/s41580-022-00566-8.

2. Schuntermann DB, *et al.*: **The central role of transfer RNAs in mistranslation**. *J Biol Chem* 2024 Sep, **300**, 107679, https://doi.org/10.1016/j.jbc.2024.107679.

3. Söll D, *et al.*: **Editorial: synthetic biology and therapeutic applications of transfer RNA**. *Front Genet* 2024 Aug, **15**, https://doi.org/10.3389/fgene.2024.1468891.

4. Deng J, *et al.*: **RNA structure determination: from 2D to 3D**. *Fundam Res* 2023 Sep 1, **3**:727–737, https://doi.org/10.1016/j.fmre.2023.06.001. Available from: https://www.sciencedirect.com/science/article/pii/S2667325823001796.

5. Ganser LR, *et al.*: **The roles of structural dynamics in the cellular functions of RNAs**. *Nat Rev Mol Cell Biol* 2019 Aug, **20**:474–489, https://doi.org/10.1038/s41580-019-0136-0. Available from: https://www.nature.com/articles/s41580-019-0136-0.

6. Lee YT, *et al.*: **The conformational space of RNase P RNA in solution**. *Nature* 2025 Jan, **637**:1244–1251, https://doi.org/10.1038/s41586-024-08336-6. Available from: https://www.nature.com/articles/s41586-024-08336-6.

7. Boniecki MJ, *et al.*: **SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction**. *Nucleic Acids Res* 2016 Apr, **44**, e63, https://doi.org/10.1093/nar/gkv1479.

8. Watkins AM, *et al.*: **FARFAR2: improved De Novo Rosetta Prediction of Complex Global RNA Folds**. *Structure* 2020 Aug, **28**:963–976.e6, https://doi.org/10.1016/j.str.2020.05.011.

9. Weigt M, *et al.*: **Identification of direct residue contacts in protein–protein interaction by message passing**. *Proc Natl Acad Sci* 2009, **106**:67–72, https://doi.org/10.1073/pnas.0805923106.

10. De Leonardis E, *et al.*: **Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction**. *Nucleic Acids Res* 2015 Dec, **43**:10444–10455, https://doi.org/10.1093/nar/gkv932.

11. Sun S, *et al.*: **RNA inter-nucleotide 3D closeness prediction by deep residual neural networks**. *Bioinformatics* 2021 May, **37**:1093–1098, https://doi.org/10.1093/bioinformatics/btaa932. 10.1093/bioinformatics/btaa932.

12. Zerihun MB, *et al.*: **CoCoNet—Boosting RNA contact prediction by convolutional neural networks**. *Nucleic Acids Res* 2021 Dec, **49**:12661–12672, https://doi.org/10.1093/nar/gkab1144.

13. Taubert O, *et al.*: **RNA contact prediction by data efficient deep learning**. *Commun Biol* 2023 Sep, **6**:913, https://doi.org/10.1038/s42003-023-05244-9.

14. Hýskova A, *et al.*: **Balancing speed and precision in protein folding: a comparison of AlphaFold2**. *ESMFold, and OmegaFold* 2025 Jun, https://doi.org/10.1101/2025.06.20.660709.

15. Bernard C, *et al.*: **State-of-the-RNArt: benchmarking current methods for RNA 3D structure prediction**. *NAR Genom Bioinformat* 2024 Jun, **6**:lqae048, https://doi.org/10.1093/nargab/lqae048.

16. Westhof E et al. The RNA-puzzles assessments of RNA-only targets in CASP16. Proteins: Struct, Funct, Bioinf. n/a. doi: 10.1002/prot.70052.

17. Lutz B, *et al.*: **Differences between cotranscriptional and free riboswitch folding**. *Nucleic Acids Res* 2014, **42**:2687–2696.

18. Bernard C, *et al.*: **RNA-TorsionBERT: leveraging language models for RNA 3D torsion angles prediction**. *Bioinformatics* 2025 Jan, **41**:btaf004, https://doi.org/10.1093/bioinformatics/btaf004.

19. Fairman CW, *et al.*: **Evaluating RNA structural flexibility: viruses lead the way**. *Viruses* 2021 Nov, **13**:2130, https://doi.org/10.3390/v13112130.

20. Fulle S, *et al.*: **Analyzing the flexibility of RNA structures by constraint counting**. *Biophys J* 2008 Jun, **94**:4202–4219, https://doi.org/10.1529/biophysj.107.113415.

21. Leontis NB, *et al.*: **Geometric nomenclature and classification of RNA base pairs**. *RNA* 2001 Jan, **7**:499–512.

22. Das J, *et al.*: **Non-canonical base pairs and higher order structures in nucleic acids: crystal structure database**

analysis. *J Biomol Struct Dyn* 2006 Oct, **24**:149–161, https://doi.org/10.1080/07391102.2006.10507108.

23. Olson WK, *et al*.: **Effects of noncanonical base pairing on RNA folding: structural context and spatial arrangements of G!!insert-eqn1/!!A pairs**. *Biochemistry* 2019 May, **58**: 2474–2487, https://doi.org/10.1021/acs.biochem.9b00122.

24. Adamczyk B, *et al*.: **RNAsolo: a repository of cleaned PDB-derived RNA 3D structures**. *Bioinformatics* 2022 Jul, **38**: 3668–3670, https://doi.org/10.1093/bioinformatics/btac386.

25. ** Szikszai M, *et al*.: **RNA3DB: a structurally-dissimilar dataset split for training and benchmarking deep learning models for RNA structure prediction**. *Journal of Molecular Biology. Computation Resources for Molecular Biology* 2024 Sep, **436**, 168552, https://doi.org/10.1016/j.jmb.2024.168552.
This is a regularly updated dataset which addresses the critical data scarcity bottleneck in RNA structure prediction by offering high-quality, non-redundant structural representatives essential for robust machine learning model training.

26. * Upadhyay U, *et al*.: **NucleoSeeker—Precision filtering of RNA databases to curate high-quality datasets**. *NAR Genom Bioinformat* 2025 Mar, **7**:lqaf021, https://doi.org/10.1093/nargab/lqaf021.
NucleoSeeker is a quality control tool that addresses the data redundancy problem plaguing RNA datasets. It provides various filters based on structural similarity, sequence similarity to remove problematic structures, significantly improving the reliability of downstream machine learning applications.

27. Leontis NB, *et al*.: **Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking**. In *RNA 3D structure analysis and prediction*. Edited by Leontis N, *et al*., Berlin, Heidelberg: Springer; 2012:281–298, https://doi.org/10.1007/978-3-642-25740-7_13.

28. Bu F, *et al*.: **RNA-puzzles round V: blind predictions of 23 RNA structures**. *Nat Methods* 2025 Feb, **22**:399–411, https://doi.org/10.1038/s41592-024-02543-9.

29. Das R, *et al*.: **Assessment of three-dimensional RNA structure prediction in CASP15. Proteins: structure**. *Funct Bioinformat* 2023, **91**:1747–1770, https://doi.org/10.1002/prot.26602.

30. Danaee P, *et al*.: **bpRNA: large-scale automated annotation and analysis of RNA secondary structure**. *Nucleic Acids Res* 2018 Jun, **46**:5381–5394, https://doi.org/10.1093/nar/gky285.

31. Tan Z, *et al*.: **TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs**. *Nucleic Acids Res* 2017 Nov, **45**:11570–11581, https://doi.org/10.1093/nar/gkx815.

32. Sloma MF, *et al*.: **Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures**. *RNA* 2016 Dec, **22**:1808–1818, https://doi.org/10.1261/rna.053694.115.

33. The RNAcentral Consortium: **RNAcentral: a hub of information for non-coding RNA sequences**. *Nucleic Acids Res* 2019 Jan, **47**:D221–D229, https://doi.org/10.1093/nar/gky1034.

34. Ontiveros-Palacios N, *et al*.: **Rfam 15: RNA families database in 2025**. *Nucleic Acids Res* 2025 Jan, **53**:D258–D267, https://doi.org/10.1093/nar/gkae1023.

35. Zhang H, *et al*.: **A new method of RNA secondary structure prediction based on convolutional neural network and dynamic programming**. *Front Genet* 2019, **10**:467.

36. Willmott D, *et al*.: **Improving RNA secondary structure prediction via state inference with deep recurrent neural networks**. *Comp and Math Biophys* 2020, **8**:36–50, https://doi.org/10.1515/cmb-2020-0002.

37. Andronescu M, *et al*.: **Efficient parameter estimation for RNA secondary structure prediction**. *Bioinformatics* 2007, **23**: i19–i28.

38. Zhao Q, *et al*.: **Review of machine learning methods for RNA secondary structure prediction**. *PLoS Comput Biol* 2021, **17**, e1009291.

39. * Takaba K, *et al*.: **Machine-learned molecular mechanics force fields from large-scale quantum chemical data**. *Chem Sci* 2024 Aug, **15**:12861–12878, https://doi.org/10.1039/D4SC00690A.
espaloma-0.3 introduces a large and well curated quantum chemical dataset as well as a graph neural network architecture to achieve an accurate Machine Learning Force Field.

40. * Seute L, *et al*.: **Grappa – a machine learned molecular mechanics force field**. *Chem Sci* 2025, **16**:2907–2930, https://doi.org/10.1039/D4SC05465B.
Grappa uses the espaloma dataset but uses an equivariant graph neural network architecture to achieve accurate simulations at little computational cost.

41. Nakamura A, *et al*.: **Fast and automated Protein-DNA/RNA macromolecular complex modeling from Cryo-EM maps**. *Briefings Bioinf* 2023 Mar, **24**:bbac632, https://doi.org/10.1093/bib/bbac632.

42. Wang X, *et al*.: **CryoREAD: de Novo Structure Modeling for Nucleic Acids in Cryo-EM Maps Using Deep Learning**. *Nat Methods* 2023 Nov, **20**:1739–1747, https://doi.org/10.1038/s41592-023-02032-5.

43. Li J, *et al*.: *DeepCryoRNA: deep learning-based RNA structure reconstruction from Cryo-EM maps*. 2025 Apr, https://doi.org/10.1101/2025.04.05.647396.

44. Yu H, *et al*.: **An interpretable RNA foundation model for exploring functional RNA motifs in plants**. *Nat Mach Intell* 2024 Dec, **6**:1616–1625, https://doi.org/10.1038/s42256-024-00946-z.

45. ** Wang N, *et al*.: **Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning**. *Nat Mach Intell* 2024 May, **6**:548–557, https://doi.org/10.1038/s42256-024-00836-4.
This method introduces a sophisticated motif-aware pre-training with multi-level masking strategies to incorporate biological priors, through innovative three-level masking of bases, subsequences, and structural motifs.

46. * Penić RJ, *et al*.: **RiNALMo: general-purpose RNA language models can generalize well on structure prediction tasks**. *Nat Commun* 2025 Jul, **16**:5671, https://doi.org/10.1038/s41467-025-60872-5.
RiNALMo is one of the largest models trained on non-coding RNA sequences. It focuses mainly on structure prediction tasks but also allows easy transfer to other downstream challenges.

47. * Upadhyay U, *et al*.: *NucleicBERT: deciphering the language of nucleic acids*. 2025 Sep, 673754, https://doi.org/10.1101/2025.09.02.
NucleicBERT extends language models to diverse RNA tasks while pioneering explainable AI integration, addressing critical interpretability challenges through advanced attention visualization and saliency methods that reveal model decision-making processes.

48. ** Shen T, *et al*.: **Accurate RNA 3D structure prediction using a language model-based deep learning approach**. *Nat Methods* 2024 Dec, **21**:2287–2298, https://doi.org/10.1038/s41592-024-02487-0.
RhoFold+ combines a language model with AlphaFold2-inspired structure module, representing one of the first end-to-end RNA structure prediction approaches. It also uses clutering and sampling from input MSAs for improved performance.

49. Baek M, *et al*.: **Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA**. *Nat Methods* 2024 Jan, **21**:117–121, https://doi.org/10.1038/s41592-023-02086-5.

50. Li Y, *et al*.: **Integrating end-to-end learning with deep geometrical potentials for Ab initio RNA structure prediction**. *Nat Commun* 2023 Sep, **14**:5745, https://doi.org/10.1038/s41467-023-41303-9.

51. Abramson J, *et al*.: **Accurate structure prediction of biomolecular interactions with AlphaFold 3**. *Nature* 2024 Jun, **630**: 493–500, https://doi.org/10.1038/s41586-024-07487-w.
** AlphaFold3 is the next-generation model of AlphaFold2. It predicts full-atom coordinates for various biomolecules, including RNA, eliminating traditional backbone-then-sidechain reconstruction approaches by using a diffusion model.

52. Wang W, *et al*.: **trRosettaRNA: automated prediction of RNA 3D structure with transformer network**. *Nat Commun* 2023 Nov, **14**:7266, https://doi.org/10.1038/s41467-023-42528-4.

53. Fu L, *et al*.: **UFold: fast and accurate RNA secondary structure prediction with deep learning**. *Nucleic Acids Res* 2022 Feb, **50**:e14, https://doi.org/10.1093/nar/gkab1074.

54. Wang Z, *et al*.: **RNADiffFold: generative RNA secondary structure prediction using discrete diffusion models**. *Briefings Bioinf* 2025 Jan, **26**:bbae618, https://doi.org/10.1093/bib/bbae618.

55. Singh J, *et al*.: **RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning**. *Nat Commun* 2019 Nov, **10**:5407, https://doi.org/10.1038/s41467-019-13395-9.

56. Tahmid MT, *et al*.: *BiRNA-BERT allows efficient RNA language modeling with adaptive tokenization*. 2024 Jul, https://doi.org/10.1101/2024.07.02.601703.

57. Romeijn L, *et al*.: *LncRNA-BERT: an RNA language model for classifying coding and long non-coding RNA*. 2025 Jan, https://doi.org/10.1101/2025.01.09.632168.

58. Chen K, *et al*.: **Self-supervised learning on millions of primary RNA sequences from 72 vertebrates improves sequence-based RNA splicing prediction**. *Briefings Bioinf* 2024 May, **25**: bbae163, https://doi.org/10.1093/bib/bbae163.

59. Selvaraju RR, *et al*.: **Grad-CAM: visual explanations from deep networks via gradient-based localization**. *Int J Comput Vis* 2020 Feb, **128**:336–359, https://doi.org/10.1007/s11263-019-01228-7. arXiv: 1610.02391 [cs].

60. Vig J, *et al*.: **BERTology meets biology: interpreting attention in protein language models**. *arXiv: 2006.15222* 2021 Mar, https://doi.org/10.48550/arXiv.2006.15222 [cs].

61. Lundberg S, *et al*.: **A unified approach to interpreting model predictions**. *arXiv: 1705.07874* 2017 Nov, https://doi.org/10.48550/arXiv.1705.07874 [cs].

62. Schneider B, *et al*.: **When will RNA get its AlphaFold moment?** *Nucleic Acids Res* 2023 Oct, **51**:9522–9532, https://doi.org/10.1093/nar/gkad726.