Towards Data-Efficient Deep Learning for RNA 3D Structure Prediction and Design

Yimeng Liu[1]

[1]University of Toronto, Department of Computer Science
[1] yoyoyimeng.liu@mail.utoronto.ca

Abstract

RNA 3D structure prediction is essential for understanding regulatory mechanisms, catalysis, and therapeutic RNA design, yet progress has lagged behind proteins due to limited structural data and the complexity of RNA folding. This work proposes a data-efficient, physics-informed deep learning framework for full atomistic prediction of transfer RNA (tRNA) tertiary structures directly from sequence. Our approach will integrate pretrained RNA embeddings, predicted secondary structure constraints, and SE(3)-equivariant graph attention to model long-range geometric relationships. A two-stage design will first predict global phosphate backbone coordinates, then reconstruct nucleobase atoms using a local geometry-aware decoder. A multi-objective loss will combine geometric accuracy with chemical and biophysical plausibility to enforce valid torsion angles, base-pairing, and steric constraints. We will benchmark against physics-based (VFold) and neural network–based (DeepFoldRNA) models to assess generalization under data scarcity. Ultimately, this framework aims to advance RNA 3D modeling with improved stability, interpretability, and capacity to generalize beyond well-characterized RNA families, supporting future applications in rational RNA engineering and structure-guided RNA design.

Background

RNA molecules play essential roles in protein synthesis, catalysis, and gene regulation. Yet, experimental determination of RNA structures remains costly and low-throughput. The resulting scarcity of high-quality structural data poses a major obstacle for data-driven modeling. As a result, although deep learning has transformed protein structure prediction, analogous progress for RNA has been limited by insufficient training data and the complexity of RNA tertiary interactions.

Early computational approaches, such as VFold (Li et al. 2022), model RNA tertiary structure using physics-based thermodynamic constraints; however, these methods rely on extensive conformational sampling and scale poorly with increasing molecular size. Recent deep learning approaches (Singh et al. 2019; Chen et al. 2020; Shen et al. 2024; He et al. 2024) have leveraged neural architectures and self-supervised learning to improve structure prediction. However, many of these models are limited to secondary structure. Those that address tertiary folding often yield poor results, as the scarcity of RNA tertiary structures (fewer than 6,000) limits training data, leading to overfitting and poor generalization. This gap motivates the development of data-efficient, physics-informed approaches. We propose a deep learning framework focused on transfer RNA (tRNA) as a tractable benchmark. tRNA combines higher data availability with complex folding challenges: while its secondary structure is conserved, its functional 3D geometry relies on precise long-range tertiary interactions (e.g., D-loop/T-loop contacts), providing a rigorous test for biophysical validity.

Problem Definition

Input: A transfer RNA (tRNA) nucleotide sequence

$$(x_1, x_2, \ldots, x_L), \quad x_i \in \{A, U, G, C\}$$

where L denotes sequence length and each x_i specifies the base identity of nucleotide i.
Output: A complete 3D tertiary structure of the tRNA represented by atomic coordinates:

- phosphorus backbone coordinates, $P_i \in R^3$,
- predicted heavy-atom coordinates for each nucleobase,
- optional annotations describing hydrogen-bond donor/acceptor groups for base-pairing geometry.

This representation supports evaluation of canonical and noncanonical base pairs, backbone geometry, and tertiary interactions such as D–loop and T–loop contacts.

Proposed Approach

Data Collection

We will leverage experimentally resolved RNA structures from:

- **RCSB PDB**: ∼1,500 tRNA-containing structures (filtered to isolate RNA chains from protein complexes) (Berman et al. 2000).
- **RNA 3D Hub**: Curated, non-redundant RNA datasets used to benchmark structural quality and identify representative conformers (Petrov, Zirbel, and Leontis 2013).

- **RNACentral**: Used to unify sequence identifiers, map metadata across databases, and retrieve homologous sequences for evolutionary analysis (The RNAcentral Consortium 2018).

Benchmark Methods

We benchmark against two complementary and widely-used approaches for RNA 3D structure prediction:

Physics-based: *VFold* (Li et al. 2022), which ranked first in the CASP16 RNA prediction category in both template-free and template-based modeling, making VFold a strong reference point (Kretsch et al. 2025).

Deep learning-based: *DeepFoldRNA* (Pearce, Omenn, and Zhang 2022), a deep learning system with clearly defined training and testing splits, enabling robust performance comparison.

These benchmarks provide insight into strengths of physics-driven vs. data-driven paradigms for RNA tertiary structure prediction.

Architecture Design

The proposed framework consists of two major components: (1) global structure prediction of the phosphate backbone, and (2) local reconstruction of RNA base atoms. The model will be driven by both learned residue embeddings and predicted secondary structure constraints.

Global Backbone Module. First, the RNA primary sequence is processed by RNA-FM (Wang et al. 2024) to obtain contextualized embeddings for each nucleotide. In parallel, secondary structure probability matrices are predicted using RNAformer (Franke et al. 2024). These form the basis of a nucleotide graph representation:

- **Nodes:** nucleotide identity (A/U/G/C/modified bases), RNA-FM embeddings
- **Edges:** backbone connectivity and predicted interaction edges from RNAformer contact probabilities

We will apply an SE(3)-Transformer (Fuchs et al. 2020) to propagate geometric information across the RNA graph. Multi-head self-attention will capture heterogeneous structural relationships, while SE(3) equivariance will ensure predictions transform faithfully under rotation and translation. The final output of this stage will consist of:

1. 3D coordinates of the phosphorus atom for each nucleotide ($P_i \in R^3$)
2. Per-nucleotide global structural embeddings H_i, encoding secondary and tertiary context

To capture diverse structural motifs, we will employ a sparsely-gated Mixture-of-Experts (MoE) in the final layers. This allows specific experts to automatically specialize in distinct topologies, such as A-form helices versus non-canonical loops, enhancing model capacity efficiently.

Local Base Reconstruction Module. To recover the full atomic structure, we refine the backbone predictions with a residue-level geometric decoder. For each nucleotide i, we extract a local neighborhood of backbone atoms (within a 10 Å radius). To ensure rotational invariance, these coordinates will be transformed into a canonical local frame before being concatenated with the global embedding H_i and base identity.

This set of local tokens will be processed by a lightweight geometry-aware transformer decoder, which predicts the relative coordinates of heavy base atoms for nucleotide i. The predicted coordinates will then be transformed back to the global coordinate system, yielding a complete atomistic RNA model.

This two-stage design would allow long-range tertiary interactions to guide backbone geometry while enforcing local chemical consistency during base reconstruction.

Loss Function

We will optimize a multi-objective geometric loss

$$L = \sum_{i=1}^{5} \lambda_i L_i,$$

where:

- λ_i: adaptive loss weights
- L_1: global heavy-atom RMSD to ground truth
- L_2: base-pair formation accuracy
- L_3: torsion and bond geometry constraints
- L_4: coarse-grained thermodynamic stability
- L_5: steric clash avoidance

This representation supports evaluation of canonical and noncanonical base pairing, stacking interactions, and tertiary motifs (Li et al. 2023).

Evaluation

Predictions will be evaluated on tRNA-containing targets within standard benchmarks such as RNA-Puzzles (Magnus et al. 2019) and CASP16 using global RMSD, base-pair interaction accuracy, and secondary–tertiary consistency. To prevent data leakage, we will ensure that none of the evaluation structures appear in the training data of any model, including our own.

To assess biophysical realism, we will additionally evaluate:

- **Energetic stability**, using folding free energy estimates ($\Delta\Delta G$) to confirm that predicted conformations are thermodynamically feasible.
- **Functional compatibility**, assessing the ribosomal span, the preservation of conserved core interactions (e.g., the Levitt pair in canonical tRNAs), and the clash score.
- **Stereochemical plausibility**, using quantum-chemical inspired principles to validate local geometry and hydrogen-bond networks.

This combined geometric and biophysical assessment will ensures that predicted tRNA structures are not only accurate in 3D but also stable and functionally meaningful, laying the groundwork for future extensions toward structure-guided RNA design.

Acknowledgments

This work was completed as part of the AAAI'26 Undergraduate Consortium. I am sincerely grateful to the Association for the Advancement of Artificial Intelligence (AAAI) for their generous support in funding my participation and for their commitment to fostering and empowering emerging researchers in AI. I also wish to express my deepest gratitude to my mentor, Happiness Eric Aigbogun, for her invaluable guidance, encouragement, and insightful feedback throughout this project. I am thankful to the University of Toronto Department of Computer Science for providing an inspiring academic environment that continues to shape my development as a researcher. Finally, I would like to express my heartfelt thanks to my parents, family, and friends for their unwavering support throughout this entire journey.

Ethics Statement

This research uses publicly available RNA sequence and structure data from established sources, including RCSB PDB, RNA 3D Hub, and RNAcentral. No human subjects, personally identifiable information, or sensitive biological materials were involved. All datasets and external computational models are appropriately cited and used in accordance with their respective licenses. The methods comply with the AAAI Ethics Code and prioritize scientific transparency, reproducibility, and responsible use of AI in biological research. The work does not introduce foreseeable risks, dual-use concerns, or ethical issues that may negatively impact human welfare or safety.

References

Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; and Bourne, P. E. 2000. The Protein Data Bank. *Nucleic Acids Research*, 28(1): 235–242.

Chen, X.; Li, Y.; Umarov, R.; Gao, X.; and Song, L. 2020. RNA Secondary Structure Prediction by Learning Unrolled Algorithms. *arXiv preprint arXiv:2002.05810*.

Franke, J. K.; Runge, F.; Köksal, R.; Matus, D.; Backofen, R.; and Hutter, F. 2024. RNAformer: A Simple yet Effective Model for Homology-Aware RNA Secondary Structure Prediction.

Fuchs, F.; Worrall, D.; Fischer, V.; and Welling, M. 2020. SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1970–1981. Curran Associates, Inc.

He, S.; Huang, R.; Townley, J.; Kretsch, R. C.; Karagianes, T. G.; Cox, D. B. T.; Blair, H.; Penzar, D.; Vyaltsev, V.; Aristova, E.; Zinkevich, A.; Bakulin, A.; Sohn, H.; Krstevski, D.; Fukui, T.; Tatematsu, F.; Uchida, Y.; Jang, D.; Lee, J. S.; and Shieh, R. 2024. Ribonanza: deep learning of RNA structure through dual crowdsourcing. *bioRxiv*.

Kretsch, R. C.; Hummer, A. M.; He, S.; Yuan, R.; Zhang, J.; Karagianes, T.; Cong, Q.; Kryshtafovych, A.; and Das, R. 2025. Assessment of Nucleic Acid Structure Prediction in CASP16. *Proteins Structure Function and Bioinformatics*.

Li, J.; Zhang, S.; Zhang, D.; and Chen, S.-J. 2022. VfoldPipeline: a web server for RNA 3D structure prediction from sequences. *Bioinformatics*, 38(16): 4042–4043.

Li, Y.; Zhang, C.; Feng, C.; Pearce, R.; Lydia Freddolino, P.; and Zhang, Y. 2023. Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction. *Nature Communications*, 14(1): 5745.

Magnus, M.; Antczak, M.; Zok, T.; Wiedemann, J.; Lukasiak, P.; Cao, Y.; Bujnicki, J. M.; Westhof, E.; Szachniuk, M.; and Miao, Z. 2019. RNA-Puzzles toolkit: a computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools. *Nucleic Acids Research*, 48(2): 576–588.

Pearce, R.; Omenn, G. S.; and Zhang, Y. 2022. De Novo RNA Tertiary Structure Prediction at Atomic Resolution Using Geometric Potentials from Deep Learning. *bioRxiv*.

Petrov, A. I.; Zirbel, C. L.; and Leontis, N. B. 2013. Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. *RNA*, 19(10): 1327–1340.

Shen, T.; Hu, Z.; Sun, S.; Liu, D.; Wong, F.; Wang, J.; Chen, J.; Wang, Y.; Hong, L.; Xiao, J.; Zheng, L.; Krishnamoorthi, T.; King, I.; Wang, S.; Yin, P.; Collins, J. J.; and Li, Y. 2024. Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nature Methods*, 21(12): 2287–2298.

Singh, J.; Hanson, J.; Paliwal, K.; and Zhou, Y. 2019. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature Communications*, 10(1).

The RNAcentral Consortium. 2018. RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Research*, 47(D1): D1250–D1251.

Wang, N.; Bian, J.; Li, Y.; Li, X.; Mumtaz, S.; Kong, L.; and Xiong, H. 2024. Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning. *Nature Machine Intelligence*, 6(5): 548–557.