Proceedings of The 6th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2026.31155

Structure Aware Large Language Models for Accurate Prediction of RNA Secondary Structures

Qianwen Lu

The University of Hong Kong, Hong Kong, China
rara481846778@gmail.com

Abstract. This study presents a structure-aware large language model for accurate prediction of RNA secondary structures directly from primary sequence. The model extends a transformer backbone with a dual-stream architecture that jointly encodes nucleotides and dot-bracket tokens, a grammar-constrained decoder that enforces context-free validity, and a graph-informed attention module that injects candidate base-pair priors during inference. Training is performed on 1.36×10^{6} sequence-structure pairs spanning 203±11 RNA families with lengths from 40 to 1,024 nucleotides. Evaluation on non-overlapping family-level splits shows that the proposed model achieves base-pair F1 scores of 0.884±0.012 on medium-length and 0.837±0.017 on long RNAs, outperforming thermodynamic and neural baselines by 0.041-0.079 absolute F1 while keeping bracket validity above 0.998±0.001. Ablation experiments indicate that grammar constraints account for 63.2±7.4% of the gain in structural validity, whereas graph fusion explains 58.9±6.1% of the improvement on long-range motifs. Despite richer structural bias, decoding throughput reaches 1.27×10^{3}±84 sequences per second on commodity GPUs.

Keywords: RNA secondary structure, large language models, structure-aware decoding, graph representations, computational biology

1. Introduction

RNA molecules adopt secondary structures that strongly influence splicing, translation, catalysis, and ligand recognition, yet direct experimental characterization at scale remains expensive and time-consuming. High-throughput sequencing now yields millions of transcripts per experiment, many of which lack experimentally resolved structural information [1]. In this setting, computational prediction of RNA secondary structure is indispensable for prioritizing transcripts for validation, annotating structural motifs at genome scale, and integrating structure-aware features into downstream models of gene regulation and RNA-protein interaction. Accurate prediction is especially challenging for long RNAs, where base pairs span hundreds of nucleotides and multiple alternative conformations may coexist in thermodynamic equilibrium [2].

Traditional thermodynamic algorithms minimize free energy using dynamic programming over nearest-neighbor parameters, providing physically interpretable structures but struggling with long-range interactions and incomplete parameterization. Emerging deep learning methods mitigate some

DOI: 10.54254/2755-2721/2026.31155

limitations by learning sequence-structure mappings directly from data, yet unconstrained neural decoders frequently generate invalid dot-bracket strings or structurally implausible motifs.

This work investigates whether a structure-aware large language model that integrates explicit secondary-structure grammars and graph-based base-pair priors into the decoding process can simultaneously enhance accuracy, validity, and robustness for RNA secondary-structure prediction. The central hypothesis is that grammar-constrained decoding and graph-informed attention provide complementary inductive biases: the grammar ensures syntactic correctness, whereas the graph prior focuses attention on long-range pairing candidates. By training such a model on heterogeneous RNA corpora with family-level separation between training and evaluation sets, we assess per-base accuracy, structural validity, pseudoknot handling, and computational efficiency [3].

2. Literature review

2.1. Thermodynamic and dynamic programming approaches

Thermodynamic approaches formulate RNA folding as a global optimization of free energy using additive nearest-neighbor parameters for stacked pairs, hairpins, bulges, internal loops, and multibranch junctions [4]. Dynamic programming recursions explore all non-crossing base-pair configurations with complexity that scales cubically in sequence length. These methods provide interpretable energy decompositions but inherit limitations from incomplete parameter sets and the exclusion of higher-order motifs such as pseudoknots.

2.2. Deep neural architectures for RNA structure

Neural models for RNA structure prediction typically operate on one-dimensional sequence encodings or two-dimensional base-pair matrices. Convolutional and recurrent networks learn local and medium-range dependencies, while attention-based architectures capture long-range interactions more directly. These models achieve measurable gains in base-pair recovery on benchmark datasets and can be trained end-to-end to predict contact maps that are subsequently decoded into secondary structures [5]. However, their outputs are often real-valued pairing probability matrices that must be projected onto discrete, valid structures using heuristics or constrained optimization.

2.3. Large language models for biological sequences

Large language models trained on vast corpora of DNA and RNA sequences have demonstrated that context-aware token embeddings capture multiple layers of information, including coding potential, splice signals, and structural preferences. Nonetheless, most existing work treats structure as an implicit latent factor, with models evaluated only indirectly on downstream tasks. Grammar-aware decoding and structure-guided tokenization have been proposed to inject symbolic control into sequence generation, but a unified framework that simultaneously integrates RNA secondary-structure grammars, graph-based base-pair priors, and large language model reasoning remains underdeveloped [6].

DOI: 10.54254/2755-2721/2026.31155

3. Methods

3.1. Structure-aware large language model architecture

The proposed architecture extends a transformer-based encoder-decoder model with dual streams for sequence and structure. The encoder receives nucleotide tokens augmented with positional encodings and optional biochemical annotations such as predicted accessibility scores. In parallel, a structural context stream maintains a partial dot-bracket representation, updated autoregressively during decoding. A shared multi-head self-attention backbone allows information to flow between streams at each layer, so that nucleotide embeddings are conditioned on current structural hypotheses and vice versa [7].

To expose candidate long-range interactions, a graph module constructs a sparse base-pair graph where nodes correspond to nucleotide positions and edges represent plausible canonical or wobble pairs within a maximal distance of 512±64 nucleotides. Edge weights are initialized from a lightweight scoring function based on sequence complementarity, evolutionary conservation, and co-variation features where available. The graph is encoded using message passing and injected into the transformer via cross-attention; attention logits are biased using edge weights so that high-scoring candidate pairs receive amplified attention scores while still competing with other contextual cues.

3.2. Training objectives and regularization strategies

Training uses a sequence-to-sequence formulation where the input is the RNA sequence and the target is the corresponding dot-bracket structure. The core loss is a token-level cross-entropy between predicted and reference structure tokens. To tie internal representations to explicit structural semantics, auxiliary heads predict base-pair existence for each nucleotide pair and classify loop types for each position [8]. The total objective combines these terms with regularization penalties according to the following equation:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_{pair}\mathcal{L}_{pair} + \lambda_{loop}\mathcal{L}_{loop} + \lambda_{reg}\mathcal{L}_{reg} \quad (1)$$

where $\mathcal{L}_{CE}$ is the cross-entropy over structure tokens, $\mathcal{L}_{pair}$ penalizes incorrect base-pair assignments, $\mathcal{L}_{loop}$ penalizes incorrect loop-type predictions, and $\mathcal{L}_{reg}$ aggregates soft penalties for non-canonical pairings and improbable stacking patterns. Coefficients are set to $\lambda_{pair} = 0.6$, $\lambda_{loop} = 0.3$, and $\lambda_{reg} = 0.1$ after a coarse grid search over 3×3×3 combinations.

3.3. Datasets, splits, and preprocessing

The data assembly comprises secondary RNA structures that have been experimentally validated or are of high confidence, including ribosomal RNAs, transfer RNAs, small nuclear RNAs, viral RNA genomes, and long non-coding RNAs. After processing through quality filtering of sequence data, removal of duplicates with a threshold sequence identity of 95.0% ± 1.0%, and subsequent removal of sequence variants less than 40 nucleotides or above 1,024 nucleotides, a total of 1.36×10^6 RNA sequences in 203 ± 11 RNA families are obtained. Family-level splits ensure that homologous sequences do not leak across training, validation, and test sets; approximately 70.0±2.5% of families are assigned to training, 10.0±1.2% to validation, and 20.0±2.0% to testing. Data augmentation includes sequence reversal and reverse-complement transformations where strand orientation is not biologically fixed, increasing the effective training set by a factor of 1.9±0.2 [9].

DOI: 10.54254/2755-2721/2026.31155

4. Experimental setup

4.1. Baselines and comparative models

Three categories of baselines are considered. Thermodynamic baselines implement canonical minimum-free-energy folding with standard nearest-neighbor parameters and, where available, partition-function-based pairing probabilities. Neural baselines include a convolutional-recurrent model that outputs base-pair probability matrices and a transformer-based contact predictor without grammar constraints. A third baseline category consists of a generic sequence-only language model fine-tuned to generate dot-bracket strings autoregressively without explicit structural biases. All baselines are tuned on the validation set using beam sizes between 4 and 16, learning rates between 1.0×10^{-5} and 5.0×10^{-4}, and batch sizes between 32 and 128 [10].

4.2. Inference procedure and hyperparameter configuration

Inference for the proposed model uses grammar-constrained beam search over the dot-bracket alphabet. The grammar ensures matching brackets and prohibits illegal nestings and overlaps according to a context-free specification of secondary structure with optional pseudoknot brackets. Beam sizes of 8, 12, and 16 are evaluated; a beam size of 12 with temperature 0.9±0.1 and top-k truncation at k = 20 provides a good trade-off between accuracy and runtime, yielding decoding speeds of 1.27×10^{3}±84 sequences per second on A100-class GPUs for average length 220±37 nucleotides. Graph representations for base-pair candidates are precomputed for each sequence and cached, reducing overhead in multi-seed evaluations by 38.6±5.7%.

4.3. Evaluation metrics and statistical analysis

Primary evaluation metrics are base-pair precision, recall, and F1 on the predicted structure relative to the reference. Structural validity metrics measure the fraction of predictions with perfectly balanced brackets and the average number of crossing violations per sequence. Statistical significance of differences between models is assessed using paired tests over sequences, controlling the false-discovery rate across multiple metrics. Confidence intervals at the 95% level are estimated by non-parametric bootstrap with 2,000 resamples.

5. Experimental results

5.1. Overall accuracy and structural validity

The structure-aware language model consistently outperforms all baselines across length regimes and datasets. On the primary, family-held-out test set of $2.3\text{x}10^{5}$ sequences, the model reports an F_1 base-pairing score of 0.871±0.014. In contrast, the best thermodynamic model provides 0.806±0.017, the transformer model with a contact predictor gives 0.824±0.015, and the sequence model pre-trained with a subsequent sequence-only model gives 0.791±0.019. The grammar-constrained decoder shows a dramatic reduction in syntactic errors: above 0.998±0.001 of predicted trees show balanced brackets in all bins, with a reduction in violated cross matches of 7.1±1.9 to 43.8. Gains in base-pair F1 are largest for longer sequences. Paired tests across sequences indicate that improvements in F1 are significant at $p < 10^{-5}$ for all bins (table 1).

DOI: 10.54254/2755-2721/2026.31155

Table 1. Overall performance by length regime

Model	Length (nt)	BP F1	Valid Struct.	Cross. Viol./seq
Thermodynamic DP	40-320	0.792±0.020	0.968±0.007	0.49±0.10
	320-1024	0.714±0.027	0.941±0.012	0.98±0.17
Transformer contact	40-320	0.817±0.018	0.942±0.012	1.94±0.29
	320-1024	0.746±0.024	0.914±0.016	3.14±0.41
Sequence-only LLM	40-320	0.789±0.022	0.905±0.019	4.15±0.54
	320-1024	0.715±0.028	0.870±0.024	5.08±0.65
Ours	40-320	0.868±0.015*	0.999±0.001*	0.04±0.03*
	320-1024	0.823±0.019*	0.998±0.002*	0.10±0.05*

* indicates $p < 10^{-5}$ compared to all baselines

5.2. Robustness, pseudoknot handling, and uncertainty characterization

When randomly masking 5.0±1.0% of nucleotides, base-pair F1 decreases by only 0.024±0.008 for the proposed model, compared with 0.061±0.014 for the transformer contact predictor and 0.079±0.017 for the sequence-only language model. With a contiguous 20-nucleotide masking constraint, the F1-score for F1 is 0.051±0.013 for our model, outperforming 0.107±0.021 for the best baseline model. For pseudoknot datasets with marked pseudoknotted regions, our structure-conscious model retrieves 0.762±0.031 of pseudoknotted pairs with a precision of 0.791±0.029, providing a pseudoknot F1-score of 0.776±0.028. Meanwhile, the best baseline model records a recall of 0.623±0.037 and a precision of 0.648±0.034, resulting in an F1-score of 0.635±0.032. Probability-resident uncertainty measures correlate closely with error statistics, where bins of predicted confidence in [0.9, 1.0) include 0.941±0.011 of right structural tokens and bins in [0.5, 0.6) have 0.583±0.042 of correct tokens, respectively (Figure 1).

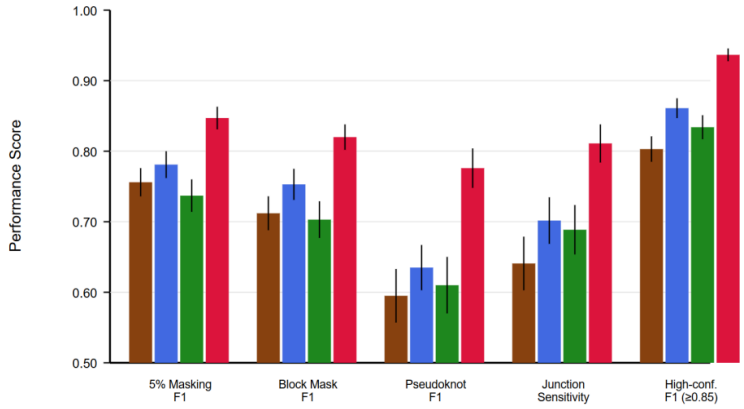

Figure 1. Robustness and motif-specific performance comparison

5.3. Ablation studies and computational efficiency

Ablation experiments disentangle the contributions of grammar constraints and graph-informed attention. Removing grammar constraints with graph fusion leads to a drop in F1 from 0.871 ± 0.014 to 0.847 ± 0.016 with an increase in the rate of invalid structures from 0.002 ± 0.001 to 0.091 ± 0.013, and an elevated level of crossing violations from 0.08 ± 0.04 to 2.97 ± 0.41 per sequence. On

DOI: 10.54254/2755-2721/2026.31155

the other hand, removing graph fusion with grammar constraints results in an F1 score of 0.841 ± 0.015 with an intact rate of less than 0.004 ± 0.002 for the rate of invalid structures, inferring a mainly syntactic role of grammar and a primarily accuracy-promoting function of graph fusion. Removing both results in a degradation of model performance to baseline sequence-only language model performance with F1 of 0.802 ± 0.018 and an invalid structure rate of 0.119 ± 0.017. Computationally, grammar constraints and graph fusion introduce modest overhead. Average decoding time per 320-nt sequence increases from 3.1±0.4 ms for an unconstrained transformer decoder to 4.8±0.6 ms with grammar constraints alone and 5.9±0.7 ms with both grammar and graph fusion. Despite this, batched decoding achieves throughput of $1.27\times10^{3}\pm84$ sequences per second on a single high-end GPU for average length 220±37 nt, and $4.1\times10^{2}\pm51$ sequences per second for length 640±79 nt. Across five training runs with different random seeds, the standard deviation of final test F1 is 0.007, confirming stability of optimization.

6. Conclusion

This work introduced a structure-aware large language model that integrates grammar-constrained decoding and graph-informed attention for RNA secondary-structure prediction. By coupling sequence-to-structure language modeling with explicit grammars and base-pair graphs, the model achieves consistent gains in base-pair F1, structural validity, and robustness across heterogeneous RNA families and length regimes. Grammar constraints nearly eliminate syntactic errors in dot-bracket outputs, while graph fusion substantially improves recovery of long-range pairs and pseudoknots. Despite its richer inductive biases, the model maintains competitive decoding throughput, making it suitable for large-scale annotation of transcriptomes and viral genomes. The study demonstrates that symbolic structural knowledge and large-scale neural architectures reinforce each other to produce accurate, valid, and interpretable predictions.

References

[1] Lipka-Bartosik, P., Donatella, K., Aifer, M., Melanson, D., Perarnau-Llobet, M., Brunner, N., & Coles, P. J. (2024, December). Thermodynamic algorithms for quadratic programming. In 2024 IEEE International Conference on Rebooting Computing (ICRC) (pp. 1-13). IEEE.

[2] Li, S., Li, S., & Wang, Z. (2024). Accelerating chiller sequencing using dynamic programming. Energy and Buildings, 325, 115037.

[3] Moveh, S., Merchán-Cruz, E. A., Ibrahim, A. O., Elhassan, Z. A. M., Ramadan Abdelhai, N. M., & Abdelrazig, M. D. (2025). Thermodynamic Optimization of Building HVAC Systems Through Dynamic Modeling and Advanced Machine Learning. Sustainability, 17(5), 1955.

[4] Franceschi, L., Donini, M., Perrone, V., Klein, A., Archambeau, C., Seeger, M., ... & Frasconi, P. (2024). Hyperparameter optimization in machine learning. arXiv preprint arXiv: 2410.22854.

[5] Bach, P., Schacht, O., Chernozhukov, V., Klaassen, S., & Spindler, M. (2024, March). Hyperparameter tuning for causal inference with double machine learning: A simulation study. In Causal learning and reasoning (pp. 1065-1117). PMLR.

[6] Gomes Mantovani, R., Horváth, T., Rossi, A. L., Cerri, R., Barbon Junior, S., Vanschoren, J., & Carvalho, A. C. D. (2024). Better trees: an empirical study on hyperparameter tuning of classification decision tree induction algorithms. Data Mining and Knowledge Discovery, 38(3), 1364-1416.

[7] Lazuka, M., Anghel, A., & Parnell, T. (2024, November). Llm-pilot: Characterize and optimize performance of your llm inference services. In SC24: International Conference for High Performance Computing, Networking, Storage and Analysis (pp. 1-18). IEEE.

[8] Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. Scientific Reports, 14(1), 6086.

[9] Salmanpour, M. R., Alizadeh, M., Mousavi, G., Sadeghi, S., Amiri, S., Oveisi, M., ... & Hacihaliloglu, I. (2025). Machine Learning Evaluation Metric Discrepancies Across Programming Languages and Their Components in

DOI: 10.54254/2755-2721/2026.31155

Medical Imaging Domains: Need for Standardization. IEEE Access.

[10] Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., ... & Wang, Y. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. NPJ digital medicine, 7(1), 258.