RNAStructFold: MSA-free RNA 3D Structure Prediction via RNA Language-Model Attention Priors

Anonymous authors
Paper under double-blind review

Abstract

Accurate RNA three-dimensional structure prediction remains challenging. Methods including RhoFold+ and RoseTTAFold2NA use multiple sequence alignments (MSAs), whose construction adds computational cost and depends on homolog availability. We introduce RNAStructFold, an MSA-free adaptation of RhoFold+ that initializes residue and pair representations from ERNIE-RNA hidden states and attention maps. Learnable layer fusion produces residue features, attention projections initialize the Evoformer pair representation, and an auxiliary head predicts global RMSD and TM-score. To reduce coevolutionary information entering through the encoder, sequences homologous to the evaluation sets are excluded from the ERNIE-RNA pretraining corpus. The model is trained on 8,242 structures spanning 16–960 nucleotides. On RNA-Puzzles and CASP15, RNAStructFold has lower mean accuracy than RhoFold+ but improves substantially on individual targets, including PZ14, PZ17, PZ18, R1116, and R1107. Reported inference speed-ups range from $2.11\times$ to $2.73\times$. Results on the curated OrphanRNA benchmark are mixed, with both improvements and substantial losses. Language-model-derived residue and pair features therefore provide a practical route to MSA-free folding and reduce alignment-search overhead, but do not yet reliably replace alignment-derived information across diverse RNAs.

1 Introduction

RNA molecules mediate catalysis, molecular recognition, and gene regulation through their three-dimensional structures. Predicting these structures from sequence requires recovering local interactions and long-range spatial organization from comparatively limited structural supervision. Community assessments, including RNA-Puzzles and CASP15, document both progress and persistent difficulties in recovering detailed RNA geometry (Cruz et al., 2012; Das et al., 2023). These challenges make RNA folding a useful setting for studying how representations learned from abundant sequence data can support prediction when experimentally resolved structures are scarce.

Many successful RNA structure predictors use multiple sequence alignments (MSAs) to obtain information about conserved and covarying positions. RhoFold+ combines alignment-derived features with RNA language-model representations, RoseTTAFold2NA (published as RoseTTAFoldNA) uses aligned sequence inputs in its three-track architecture, and NuFold incorporates MSAs into an end-to-end folding architecture (Shen et al., 2024; Baek et al., 2024; Kagaya et al., 2025). These methods use evolutionary information in their prediction pipelines. However, constructing an alignment adds a database-search stage, and the information it provides depends on the homologous sequences that can be retrieved. A predictor that builds its inputs directly from the query sequence would remove this dependency, making its computational workflow applicable even when informative alignments are unavailable.

Sequence pretraining offers a possible source of the missing structural information. ESMFold demonstrated that a protein language model can support structure prediction without explicit alignments at inference (Lin et al., 2023), and recent work has established language-model-based, single-sequence prediction for RNA (Li et al., 2026). ERNIE-RNA is particularly relevant because it uses a base-pairing-informed attention mechanism and produces attention maps with demonstrated struc-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tural signal (Yin et al., 2025). Such signal motivates using attention-derived features in a tertiary-structure predictor, but success on base-pair prediction alone does not establish that these features can replace alignment-derived inputs to a folding network.

We investigate this question with **RNAStructFold**, an MSA-free adaptation of RhoFold+ that constructs its input representations from ERNIE-RNA. The adaptation retains the downstream geometric reasoning and coordinate-refinement pipeline, providing a common architectural reference for evaluating the resulting system. Our contributions are threefold:

- **An attention-derived input representation.** We combine ERNIE-RNA hidden states through learnable layer fusion and project its attention maps into pair features, supplying both residue and pair representations without a query-specific MSA.
- **An integrated single-sequence prediction system.** We train the adapted model on a curated dataset spanning 16–960 nucleotides and include an auxiliary head that predicts global RMSD and TM-score without reference coordinates at inference.
- **An evaluation of accuracy and computational cost.** We compare the system with RhoFold+ on RNA-Puzzles and CASP15 targets, supplemented by the curated OrphanRNA benchmark, and examine variation across targets and sequence lengths.

The results reveal an accuracy–runtime trade-off. Across the three benchmark sets, our evaluation records a speed-up of approximately 2.1–$2.7\times$ relative to RhoFold+. Mean TM-score remains lower on each set, although individual targets show improved structural agreement. The largest mean TM-score gap occurs on OrphanRNA, identifying low-homology prediction as an unresolved limitation in this evaluation. These findings characterize both the practical benefit and the current limitations of using ERNIE-RNA representations as inputs to a single-sequence folding system.

2 Related Work

RNA tertiary-structure prediction. Learning-based RNA folding methods differ in both the information supplied to the model and the mechanism used to construct coordinates. DeepFoldRNA predicts geometric restraints and converts them into atomic structures through optimization (Pearce et al., 2022). DRfold combines predicted local frames with learned geometric potentials for structure assembly and refinement (Li et al., 2023). RhoFold+ integrates RNA-FM representations and MSA-derived information into a geometric folding pipeline (Shen et al., 2024). RoseTTAFold2NA, described as RoseTTAFoldNA in the associated publication, extends a three-track architecture to nucleic acids and protein–nucleic-acid complexes and uses aligned sequence inputs (Baek et al., 2024). NuFold uses an end-to-end architecture with a flexible nucleobase representation to model nucleotide geometry (Kagaya et al., 2025). These approaches provide the geometric modeling components on which our adaptation builds.

Representations learned from RNA sequences. RNA-FM demonstrated that pretraining on unlabeled RNA sequences can support multiple structural and functional prediction tasks (Chen et al., 2022). RiNALMo further demonstrated transfer to structure-related tasks, including secondary-structure prediction across RNA families (Penić et al., 2025). RNA-MSM takes a different approach: it learns from MSAs and produces embeddings and attention maps informative about RNA structure (Zhang et al., 2024). It therefore provides evidence for structural information in attention while retaining an explicit alignment requirement. ERNIE-RNA instead introduces a base-pairing-informed attention mechanism and evaluates its representations on structural and functional tasks (Yin et al., 2025). Our work uses its hidden states and attention maps as complementary inputs to a downstream folding network.

Single-sequence structure prediction. ESMFold established an influential precedent for mapping pretrained sequence representations to protein coordinates without constructing an MSA for each target (Lin et al., 2023). Single-sequence RNA folding also predates this work: DRfold learns from query sequences and predicted secondary structures (Li et al., 2023), while DRfold2 combines a pretrained RNA Composite Language Model with a denoising structure module (Li et al., 2026). More recently, the NUMonomer preprint presented nucleic-acid structure prediction from primary sequence alone (Si et al., 2026). RNAStructFold belongs to this existing line of research; its specific

focus is the use of ERNIE-RNA layer representations and attention-derived pair features within a RhoFold+ folding backbone.

Integrating language models with folding architectures. The effectiveness of pretrained features depends on how they interact with the structure predictor. Zou et al. (2025) studied RNA language-model fusion in an AlphaFold 3-style architecture and found substantial variation across integration strategies, with their strongest results obtained when language-model features were combined with MSAs. This finding cautions against treating language-model representations and alignment-derived information as universally interchangeable. RNAStructFold examines a particular alternative: constructing the initial single and pair representations from ERNIE-RNA and evaluating the resulting MSA-free system against its RhoFold+ reference. Here, MSA-free refers to the absence of a query-specific alignment input; it does not imply that sequence pretraining contains no evolutionary information.

3 METHOD

RNAStructFold maps an RNA sequence $x = (x_1, \ldots, x_L)$ to atomic coordinates without constructing a query-specific MSA. The model replaces the MSA/RNA-FM input branch of RhoFold+ with ERNIE-RNA features, while retaining its downstream folding and coordinate-refinement modules. This design preserves a common geometric architecture; the resulting comparison is between trained systems, since the representation change is accompanied by an expanded training dataset.

3.1 RESIDUE AND PAIR REPRESENTATIONS

Language-model features. For a sequence of length L, ERNIE-RNA produces contextual hidden states $H^{(\ell)} \in \mathbb{R}^{L\times d}$ at transformer layer ℓ, together with multi-head self-attention maps. We collect the hidden states from N layers and concatenate the attention channels across layers and heads into $A \in \mathbb{R}^{L\times L\times C}$. These features provide two complementary views of the sequence: contextual information at each nucleotide and learned relationships between nucleotide pairs. Attention values are used as input features, rather than interpreted as physical distances or calibrated contact probabilities.

Learnable layer fusion. A learned scalar α_ℓ controls the contribution of each layer. Softmax-normalized weights define the fused residue representation, which is projected into the single-representation dimension d_s of the folding network:

$$w_\ell = \frac{\exp(\alpha_\ell)}{\sum_{j=1}^{N}\exp(\alpha_j)}, \qquad H_{\text{fused}} = \sum_{\ell=1}^{N} w_\ell H^{(\ell)}, \qquad S_0 = H_{\text{fused}} W_s + b_s. \tag{1}$$

The fusion allows supervision from the folding task to select among pretrained layers instead of using only the final hidden state.

Attention-derived pair initialization. The attention tensor is transformed by a two-dimensional convolution, layer normalization, and a linear projection:

$$Z' = \text{LayerNorm}(\text{Conv2D}(A)), \qquad Z_0 = \text{Linear}_p(Z') \in \mathbb{R}^{L\times L\times d_z}. \tag{2}$$

The convolution combines attention features on the nucleotide-pair grid, and the projection matches the pair dimension d_z expected by the folding trunk. The resulting S_0 and Z_0 replace alignment-derived inputs while preserving the downstream representation interfaces. This construction remains available when homolog retrieval yields few informative sequences.

3.2 GEOMETRIC FOLDING AND RECYCLING

The single and pair representations are refined by the retained RhoFold+ folding trunk, referred to as the Evoformer here. Attention, triangular operations, and transition layers update the representations before coordinate prediction. Writing the stack of K blocks as $\mathcal{E}$ gives

$$(S_K, Z_K) = \mathcal{E}(S_0, Z_0). \tag{3}$$

Algorithm 1 RNAStructFold inference

Require: RNA sequence x; number of recycling iterations $R = 4$
Ensure: Atomic coordinates X; predicted global quality scores $\widehat{r}, \widehat{t}$

$\{H^{(\ell)}\}_{\ell=1}^{N}, A \leftarrow \text{ERNIE-RNA}(x)$
$S_0 \leftarrow \text{SingleProjection}(\text{LayerFusion}(H))$
$Z_0 \leftarrow \text{PairProjection}(\text{LayerNorm}(\text{Conv2D}(A)))$
Initialize recycling states
for $r = 1, \ldots, R$ **do**
    $(S, Z) \leftarrow \mathcal{E}(S_0, Z_0)$
    $F \leftarrow \text{IPA}(S, Z)$
    $X \leftarrow \text{RNAConverter}(F)$
    $X \leftarrow \text{EGNNRefine}(X)$
    $(S_0, Z_0) \leftarrow \text{Recycle}(S, Z, X)$
end for
$(\widehat{r}, \widehat{t}) \leftarrow \text{ConfidenceHead}(S)$
return $X, \widehat{r}, \widehat{t}$

An invariant point attention (IPA) structure module predicts nucleotide frames from the refined features. The RNA coordinate converter maps these frames to atomic coordinates, and an E(3)-equivariant graph neural network (EGNN) performs coordinate refinement. These geometric components follow the RhoFold+ pipeline (Shen et al., 2024); the proposed modification concerns the representations supplied to that pipeline.

Recycling feeds the predicted structure and refined features back into the network. Four iterations are used in training and inference. Algorithm 1 summarizes the forward computation; the notation Recycle denotes the retained recycling operation.

3.3 Auxiliary quality prediction

An auxiliary head predicts the global RMSD and TM-score of a structure from the final single representation. Mean pooling over nucleotide positions yields a fixed-dimensional sequence representation, followed by a multilayer perceptron:

$$(\widehat{r}, \widehat{t}) = \text{MLP}\left(\frac{1}{L}\sum_{i=1}^{L} S_{K,i}\right). \tag{4}$$

The head is supervised by regression to reference-based structural scores computed during training. It requires no reference coordinates at inference. We describe this head as part of the architecture, but do not claim calibrated confidence or improved model selection: neither calibration nor ranking performance is evaluated in the reported experiments.

4 Experimental setup

4.1 Training data and optimization

We start from a frozen training split of 1,916 representative structures derived from the BGSU RNA 3D Hub NR 4.0 collection. Each training representative is expanded to include the corresponding cluster members. Entries already assigned to the frozen validation or test sets are excluded. Quality control removes structures with missing atomic coordinates, corrupted coordinate files, or abnormal radii of gyration, and retains sequences between 16 and 960 nucleotides. The expanded training set is screened against the frozen validation and test sets using CD-HIT-EST-2D at an 80% sequence-identity threshold; no overlaps are detected by this screen. Separately, sequences homologous to the evaluation sets are excluded from the ERNIE-RNA pretraining corpus before encoder pretraining. These screens reduce direct sequence-level leakage but do not establish the absence of more remote family-level similarity.

The curated collection contains 8,242 structures, of which 6,829 are single-chain entries. Multi-chain entries are also retained in the training collection, while the evaluation in this paper concerns

single-chain prediction. Approximately two-thirds of the collection lies between 61 and 200 nucleotides. The original representative-only split had a median length of 57 nucleotides. Expanding cluster membership therefore increases the number of available training structures and changes their length distribution, but should not be equated with a fourfold increase in independent RNA families.

Optimization uses AdamW with linear learning-rate warmup followed by cosine decay. Training processes one target per forward pass and uses gradient accumulation. CUDA bfloat16 mixed precision is used following instability observed with fp16; gradients are clipped to a global norm of 1.0, and exponential moving averages of parameters are maintained. The folding trunk and structure module use separate learning rates, with the latter assigned a lower rate.

4.2 Benchmarks and comparison protocol

Evaluation sets. The paired score table contains 47 evaluation records: 22 from RNA-Puzzles, 13 from CASP15, and 12 from a curated OrphanRNA set. RNA-Puzzles and CASP15 provide established community evaluation tasks (Cruz et al., 2012; Das et al., 2023). We use these names to identify the evaluated subsets; these experiments should not be interpreted as prospective blind participation in either assessment.

OrphanRNA is intended to probe RNAs with limited known homology and includes long and regulatory RNAs with experimentally resolved structures. Candidate families were drawn from Rfam release 15.1 and screened using homolog counts, taxonomic diversity, and covariance-model searches against RNAcentral. Only candidates with PDB structures were retained. The selection includes long RNAs between 300 and 900 nucleotides, alongside regulatory elements, internal ribosome entry sites, and aptamers. The complete set spans a broader length range. Because per-target homology statistics and numerical selection thresholds are not available in the current record, "OrphanRNA" denotes this curated benchmark and is not evidence that every target lacks informative evolutionary signal.

Metrics and aggregation. We compare RNAStructFold with RhoFold+ using the supplied paired TM-scores and coordinate RMSDs, where higher TM-score and lower RMSD are better. Means are calculated over records within each set; the overall accuracy means pool all 47 records. Local distance difference test (lDDT) scores in the accompanying table provide a secondary measure of local structural agreement. The interpretation of these metrics depends on atom selection, residue matching, and TM-score normalization; incomplete specification of these settings is a reproducibility limitation (Section 6).

For the categorical summaries, a TM-score difference within ± 0.02 is labeled *comparable*. The OrphanRNA RMSD categories use a ± 1 Å band. Both bands reproduce the tabulated counts and are descriptive conventions, not statistical equivalence tests. A "win" in these tables therefore means an improvement larger than the stated band, rather than any positive numerical difference. No significance claim is based on these counts.

The CASP15 table contains records named R1117 and R1117v2 with identical lengths and identical scores for both models. We retain both to preserve the reported benchmark totals, but do not assume that the 47 records represent 47 independent targets. Additional CASP16 visualizations are provided in Appendix A and are excluded from the three-set quantitative summary because their target-level score table is unavailable.

5 Results

5.1 Accuracy on RNA-Puzzles and CASP15

Table 1 summarizes the standard benchmark subsets. RNAStructFold has lower mean TM-score and higher mean RMSD than RhoFold+ on both. The paired observations in Figure 1 show substantial variation across targets, including improvements that are not captured by the aggregate means. This variation motivates examining individual examples alongside the overall deficit.

RNA-Puzzles. The mean TM-score is 0.226 for RNAStructFold and 0.240 for RhoFold+, an absolute difference of 0.014. Mean RMSD is 12.93 versus 11.28 Å. Several targets nevertheless show

Table 1: Performance on the standard benchmark subsets. RSF denotes RNAStructFold. TM-score is higher-is-better (↑); RMSD in Å is lower-is-better (↓). W/C/L counts are relative to RhoFold+ using the ±0.02 TM-score band.

	TM-score ↑		RMSD (Å) ↓		TM-score
Benchmark	RSF	RhoFold+	RSF	RhoFold+	W/C/L
RNA-Puzzles	0.226	0.240	12.93	11.28	5/5/12
CASP15	0.100	0.134	24.78	23.01	3/5/5

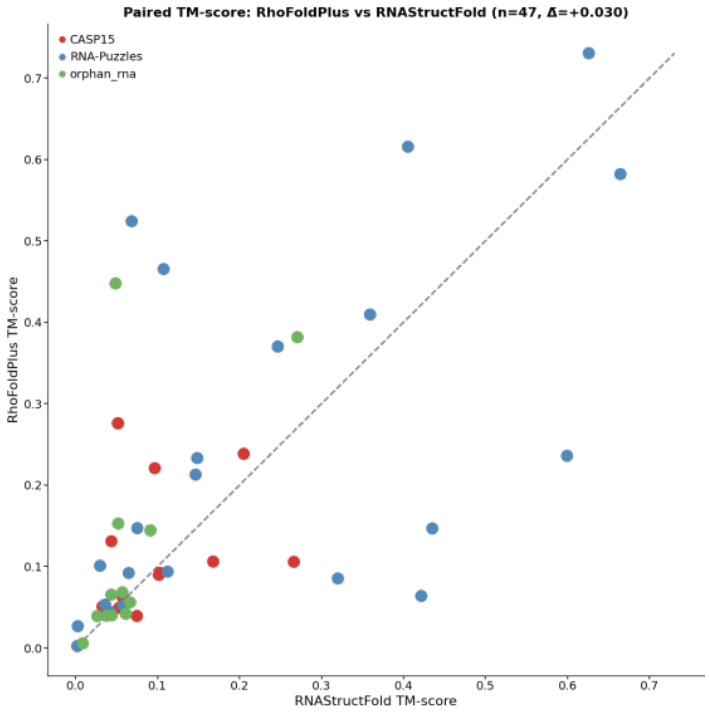


Figure 1: Paired TM-scores for the 47 evaluation records from RNA-Puzzles, CASP15, and OrphanRNA. Each point compares RhoFold+ with RNAStructFold for one record; the diagonal marks equal scores. Points on opposite sides of the diagonal indicate target-dependent gains and losses.

clear improvements on both metrics. For PZ17, TM-score increases from 0.236 to 0.600 and RMSD decreases from 6.92 to 2.54 Å. PZ18 improves from 0.085 to 0.320 in TM-score, with RMSD falling from 13.91 to 5.23 Å. The free and bound forms of PZ14 also improve in both metrics: RMSD decreases from 16.0 to 4.1 Å for PZ14 (free) and from 10.1 to 3.7 Å for PZ14 (bound). PZ3 improves from 0.582 to 0.665 in TM-score.

These gains coexist with substantial failures. For PZ5, RMSD increases from 5.77 to 28.15 Å, and for PZ7 it increases from 6.78 to 25.86 Å. Under the descriptive TM-score band, RNAStructFold records five wins, five comparable cases, and twelve losses. Thus, a small difference between the mean TM-scores does not imply consistent agreement across targets, and the losses cannot be dismissed as isolated exceptions to uniformly preserved accuracy.

CASP15. Mean TM-score is 0.100 for RNAStructFold versus 0.134 for RhoFold+, and mean RMSD is 24.78 versus 23.01 Å. The corresponding TM-score categories are three wins, five comparable cases, and five losses. R1116 is a favorable example: TM-score increases from 0.106 to 0.266, while RMSD decreases from 20.99 to 15.47 Å. R1107 also improves in TM-score, from approximately 0.106 to 0.168. Nevertheless, the low aggregate scores of both methods and the mean deficit

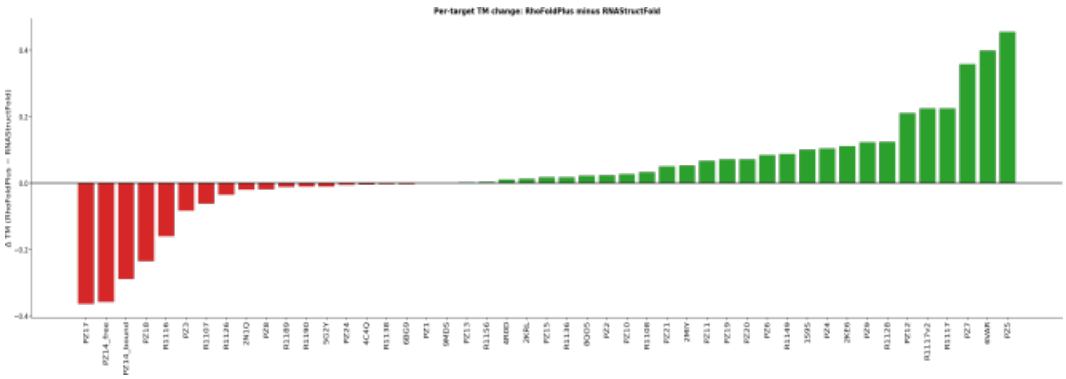


Figure 2: Per-record TM-score differences across all three benchmark sets, sorted by magnitude. The plotted difference is RhoFold+ minus RNAStructFold: negative values favor RNAStructFold and positive values favor RhoFold+. The comparison pools RNA-Puzzles, CASP15, and OrphanRNA.

Table 2: Performance on OrphanRNA. W/C/L denotes wins, comparable cases, and losses of RNAStructFold relative to RhoFold+. The comparison bands are ± 0.02 for TM-score and ± 1 Å for RMSD.

Metric	RNAStructFold	RhoFold+
Mean TM-score ↑	0.067	0.124
Mean RMSD (Å) ↓	32.00	30.76
TM-score W/C/L	0/7/5	—
RMSD W/C/L	3/3/6	—

for RNAStructFold indicate that this subset remains challenging. The aggregate comparison is also sensitive to the duplicated-score R1117 records described in Section 4.2.

5.2 Performance on OrphanRNA

OrphanRNA has the largest mean TM-score deficit: 0.067 for RNAStructFold compared with 0.124 for RhoFold+ (Table 2). Seven of twelve records fall within the TM-score comparison band, but none exceeds the baseline by more than 0.02. Mean RMSD is 32.00 versus 30.76 Å, with three wins, three comparable cases, and six losses under the RMSD band. Similarity between low scores on some targets should not be interpreted as successful recovery of their tertiary structures.

The long RNAs 9MDS and 5G2Y show lower RMSD for RNAStructFold, whereas the medium-length target 6VAR shows a large loss in both TM-score and RMSD. These examples illustrate heterogeneity within the set; they do not establish a general advantage on long or low-homology RNA. In particular, the mean lDDT is also lower for RNAStructFold, 0.414 versus 0.473, so the local agreement metric does not reverse the conclusion from TM-score. Eliminating MSA construction makes the prediction workflow applicable without homolog retrieval, but the present benchmark does not demonstrate improved accuracy in the intended low-homology regime.

5.3 Sequence length and metric dependence

Figure 3 reports the length-stratified TM-score distribution for RNAStructFold across the three sets. The 51–100 nucleotide bin is the only bin in which the model has both higher mean TM-score (0.255 versus 0.205) and lower mean RMSD (10.17 versus 10.93 Å) than RhoFold+. It contains several of the favorable RNA-Puzzles examples, including PZ17, PZ18, and both PZ14 conformations. This pattern identifies a promising subset within the present evaluation, but the small, compositionally different bins do not isolate sequence length as the cause of the advantage.

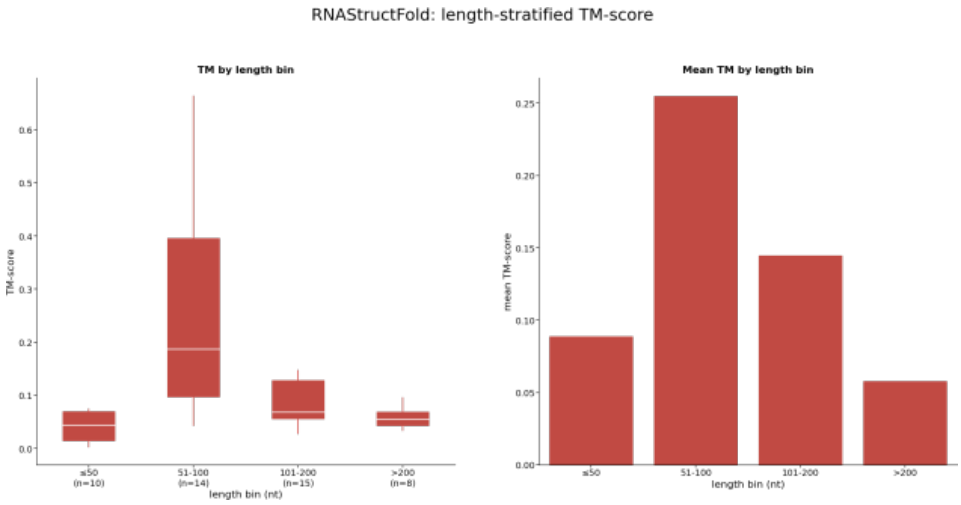


Figure 3: RNAStructFold TM-score by sequence length across the 47 evaluation records: distributions (left) and means (right). The bins contain 10, 14, 15, and 8 records, respectively. The evaluation pools RNA-Puzzles, CASP15, and OrphanRNA. Dataset composition differs across bins, so the figure is descriptive rather than a controlled test of length generalization.

Table 3: Overall performance summary for RNAStructFold, using the reported values. Accuracy means pool the listed evaluation records. Runtime ratios are relative to RhoFold+; the overall speed-up is reported separately because its aggregation rule is unavailable.

Benchmark	n	Mean TM-score ↑	Mean RMSD (Å) ↓	Speed-up
RNA-Puzzles	22	0.226	12.93	2.11×
CASP15	13	0.100	24.78	2.53×
OrphanRNA	12	0.067	32.00	2.73×
Overall	47	0.151	21.08	2.54×

For sequences longer than 200 nucleotides, mean RMSD favors RNAStructFold (44.73 versus 46.95 Å), whereas mean TM-score remains lower (0.058 versus 0.070). The corresponding mean lDDT is 0.269 versus 0.328. These large coordinate errors and lower scores do not support a broad claim of accurate long-RNA prediction. The largest deficits in mean TM-score occur in the shortest bin and the 101–200 nucleotide bin. Training on sequences up to 960 nucleotides therefore expands the supported training range, but does not establish reliable accuracy throughout that range; the longest evaluated record contains 720 nucleotides.

5.4 Runtime and aggregate performance

The reported inference speed-ups relative to RhoFold+ are 2.11× on RNA-Puzzles, 2.53× on CASP15, and 2.73× on OrphanRNA (Table 3). Removing query-specific alignment construction eliminates one stage of the prediction workflow. However, the current record does not include raw timing measurements, hardware specifications, caching conditions, or a breakdown of retrieval and network execution. The reported ratios therefore describe the evaluated implementation and should not be generalized to other hardware or interpreted as an isolated measurement of the cost of MSA generation. The overall 2.54× value is retained as reported; its aggregation rule is not specified.

Across all 47 records, mean TM-score is 0.151 for RNAStructFold versus 0.181 for RhoFold+, mean RMSD is 21.08 versus 19.50 Å, and mean lDDT is 0.477 versus 0.509. The TM-score band yields eight wins, seventeen comparable cases, and twenty-two losses (Figure 2). The descriptive comparison band does not establish statistical equivalence. Taken together, the metrics support a faster reported workflow with target-dependent accuracy gains and an overall accuracy deficit.

6 DISCUSSION AND LIMITATIONS

What the representation change establishes. RNAStructFold demonstrates how ERNIE-RNA hidden states and attention maps can supply the two input interfaces of a geometric RNA folding system. The favorable RNA-Puzzles examples show that the resulting single-sequence system can improve both TM-score and RMSD for particular targets. This is useful evidence for the integration strategy. It does not show that attention maps recover all information available from alignments, or that they encode the tertiary constraints responsible for the observed gains. The consistently lower aggregate scores, especially on OrphanRNA, define the limits of that interpretation.

What the comparison does not isolate. Retaining the geometric architecture makes the systems related, but does not provide a controlled ablation of evolutionary information. The language-model source, input projections, training-set composition, and optimization change together. Attribution requires training matched variants with the same data and budget: residue embeddings alone, attention-derived pair features alone, their combination, and a suitable alignment-based reference. Comparisons with other single-sequence predictors would also be needed to establish competitiveness within the broader MSA-free literature. The present experiments support a system-level comparison with RhoFold+, not a state-of-the-art claim.

Evaluation and reproducibility limits. The evaluation is small, the length bins mix RNA classes, and the relationship between the identical R1117 records remains unresolved. The training-set screen and ERNIE-RNA pretraining-corpus filter reduce direct sequence-level overlap but do not exclude remote family-level similarity; the pretraining filter additionally requires its search procedure and homology threshold to be reported for full reproducibility. OrphanRNA needs an explicit target manifest with homology measurements to support stronger conclusions about evolutionary scarcity. The auxiliary quality head requires calibration and ranking evaluation before it can be used as evidence of reliable confidence estimation.

Full reproduction additionally requires the complete training configuration, the policy for adapting the pretrained language model, and the treatment of multi-chain training entries. Scoring scripts must define atom selection, residue correspondence, and metric normalization. Raw runtimes and a hardware-specific timing protocol are needed to reproduce the computational comparison. These reporting gaps limit the strength of conclusions available from the current experiment record.

Scope and next steps. The current evaluation concerns single-chain RNA. Multi-chain RNA assemblies and RNA–protein complexes are outside its demonstrated scope. The most informative next experiments are matched-data representation ablations, evaluation stratified by measured homology and RNA family, and an end-to-end runtime breakdown. These would test whether the observed benefits arise from pretrained features, the expanded training collection, or implementation choices. Improving attention and memory efficiency may reduce cost further, but does not address the accuracy limitations by itself.

7 CONCLUSION

RNAStructFold integrates ERNIE-RNA residue embeddings and attention-derived pair features into the RhoFold+ geometric folding pipeline, enabling inference without a query-specific MSA. The current evaluation shows reported speed-ups of approximately 2.1–$2.7\times$, improvements on selected targets, and lower average accuracy than RhoFold+ on each benchmark set. Its contribution is a concrete integration strategy and an empirical characterization of its trade-offs. Achieving reliable gains on low-homology and long RNAs remains an open challenge that requires stronger representations and more controlled evaluation.

AI USE STATEMENT

Generative AI (OpenAI Codex) was used only to assist with manuscript restructuring and language editing. The authors reviewed the AI-assisted text and take responsibility for the final content of the paper.

ETHICS STATEMENT

This work studies computational RNA structure prediction using existing sequence and structural resources. Prediction errors can affect downstream biological interpretation, and the limited coverage of RNA classes constrains generalization. Predicted structures should be treated as hypotheses requiring independent validation; the reported benchmark limitations should accompany their use.

REPRODUCIBILITY STATEMENT

Section 3 specifies the representation construction and inference pipeline; Section 4.1 describes data curation and the available optimization settings. The source package accompanying this draft includes the 47-record paired score table, the preserved figure assets, and an analysis script that checks the reported accuracy summaries. Complete replication of training, scoring, and runtime measurements requires the additional configuration and evaluation artifacts identified in Section 6. No public code or checkpoint release is claimed in this version.

REFERENCES

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, 2024. doi: 10.1038/s41592-023-02086-5. URL https://doi.org/10.1038/s41592-023-02086-5.

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, Irwin King, and Yu Li. Interpretable RNA Foundation Model from Unannotated Data for Highly Accurate RNA Structure and Function Predictions. *bioRxiv*, 2022. doi: 10.1101/2022.08.06.503062. URL https://doi.org/10.1101/2022.08.06.503062. Preprint.

José Almeida Cruz, Marc-Frédérick Blanchet, Michal Boniecki, Janusz M. Bujnicki, Shi-Jie Chen, Song Cao, Rhiju Das, Feng Ding, Nikolay V. Dokholyan, Samuel Coulbourn Flores, Lili Huang, Christopher A. Lavender, Véronique Lisi, François Major, Katarzyna Mikolajczak, Dinshaw J. Patel, Anna Philips, Tomasz Puton, John Santalucia, Fredrick Sijenyi, Thomas Hermann, Kristian Rother, Magdalena Rother, Alexander Serganov, Marcin Skorupski, Tomasz Soltysinski, Parin Sripakdeevong, Irina Tuszynska, Kevin M. Weeks, Christina Waldsich, Michael Wildauer, Neocles B. Leontis, and Eric Westhof. RNA-Puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction. *RNA*, 18(4):610–625, 2012. doi: 10.1261/rna.031054.111. URL https://doi.org/10.1261/rna.031054.111.

Rhiju Das, Rachael C. Kretsch, Adam J. Simpkin, Thomas Mulvaney, Phillip Pham, Ramya Rangan, Fan Bu, Ronan M. Keegan, Maya Topf, Daniel J. Rigden, Zhichao Miao, and Eric Westhof. Assessment of three-dimensional RNA structure prediction in CASP15. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1747–1770, 2023. doi: 10.1002/prot.26602. URL https://doi.org/10.1002/prot.26602.

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. doi: 10.1038/s41467-025-56261-7. URL https://doi.org/10.1038/s41467-025-56261-7.

Yang Li, Chengxin Zhang, Chenjie Feng, Robin Pearce, Patricia 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, 2023. doi: 10.1038/s41467-023-41303-9. URL https://doi.org/10.1038/s41467-023-41303-9.

Yang Li, Chenjie Feng, Xi Zhang, Sho Tsukiyama, Duanyu Feng, and Yang Zhang. DRfold2 is a deep learning-based tool that enables efficient and accurate RNA structure prediction. *PLOS Biology*, 24(2):e3003659, 2026. doi: 10.1371/journal.pbio.3003659. URL https://doi.org/10.1371/journal.pbio.3003659.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://doi.org/10.1126/science.ade2574.

Robin Pearce, Gilbert S. Omenn, and Yang Zhang. De Novo RNA Tertiary Structure Prediction at Atomic Resolution Using Geometric Potentials from Deep Learning. *bioRxiv*, 2022. doi: 10.1101/2022.05.15.491755. URL https://doi.org/10.1101/2022.05.15.491755. Preprint.

Rafael Josip Penić, Tin Vlašić, Roland G. Huber, Yue Wan, and Mile Šikić. RiNALMo: general-purpose RNA language models can generalize well on structure prediction tasks. *Nature Communications*, 16(1):5671, 2025. doi: 10.1038/s41467-025-60872-5. URL https://doi.org/10.1038/s41467-025-60872-5.

Tao Shen, Zhihang Hu, Siqi Sun, Di Liu, Felix Wong, Jiuming Wang, Jiayang Chen, Yixuan Wang, Liang Hong, Jin Xiao, Liangzhen Zheng, Tejas Krishnamoorthi, Irwin King, Sheng Wang, Peng Yin, James J. Collins, and Yu Li. Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nature Methods*, 21(12):2287–2298, 2024. doi: 10.1038/s41592-024-02487-0. URL https://doi.org/10.1038/s41592-024-02487-0.

Yunda Si, Suqi Zhang, and Luonan Chen. NUMonomer enables accurate and scalable nucleic acid structure prediction from primary sequence alone. *bioRxiv*, 2026. doi: 10.64898/2026.07.20.739453. URL https://doi.org/10.64898/2026.07.20.739453. Preprint.

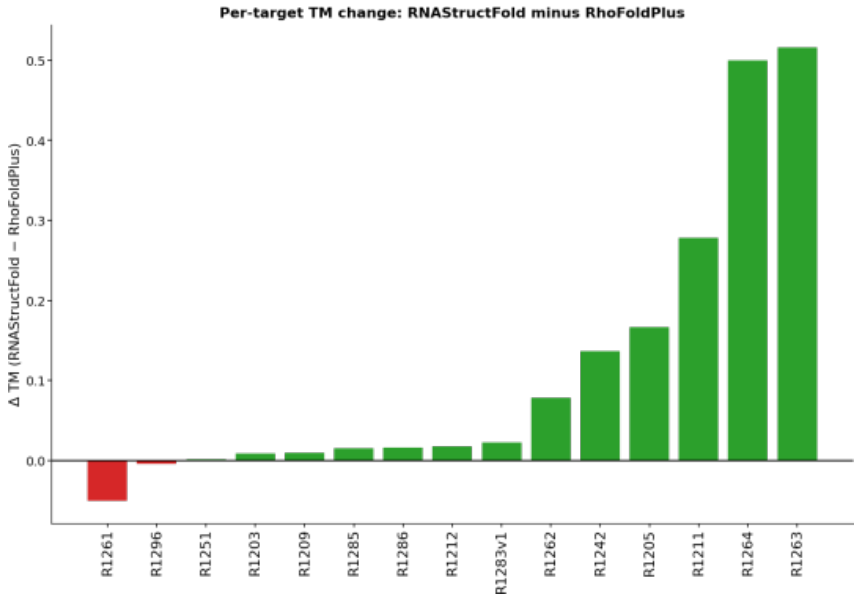
Weijie Yin, Zhaoyu Zhang, Shuo Zhang, Liang He, Ruiyang Zhang, Rui Jiang, Gan Liu, Jingyi Wang, Xuegong Zhang, Tao Qin, and Zhen Xie. ERNIE-RNA: an RNA language model with structure-enhanced representations. *Nature Communications*, 16(1):10076, 2025. doi: 10.1038/s41467-025-64972-0. URL https://doi.org/10.1038/s41467-025-64972-0.

Yikun Zhang, Mei Lang, Jiuhong Jiang, Zhiqiang Gao, Fan Xu, Thomas Litfin, Ke Chen, Jaswinder Singh, Xiansong Huang, Guoli Song, Yonghong Tian, Jian Zhan, Jie Chen, and Yaoqi Zhou. Multiple sequence alignment-based RNA language model and its application to structural inference. *Nucleic Acids Research*, 52(1):e3, 2024. doi: 10.1093/nar/gkad1031. URL https://doi.org/10.1093/nar/gkad1031.

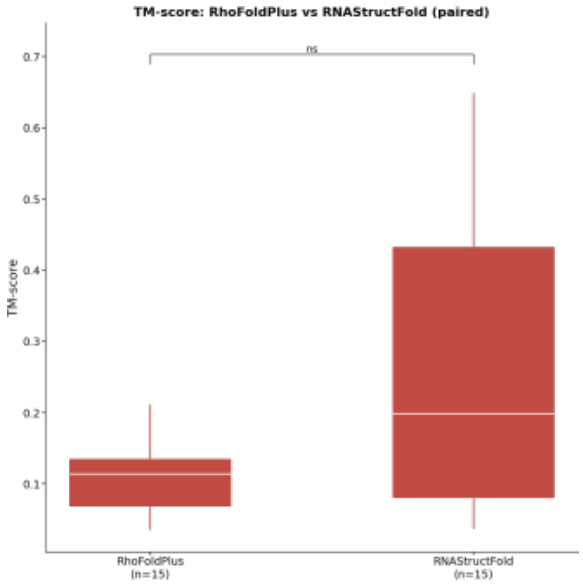
Shuxian Zou, Jiayou Zhang, Bingkang Zhao, Hui Li, Eric P. Xing, and Le Song. Improving RNA 3D Structure Prediction via Language Model-Augmented AlphaFold 3. In *NeurIPS 2025 Workshop on Machine Learning for Structural Biology*, 2025. URL https://www.mlsb.io/papers_2025/72.pdf.

A ADDITIONAL CASP16 VISUALIZATIONS

The CASP16 visualizations in Figure 4 concern a separate evaluation from the 47-record benchmark. The corresponding score table and evaluation protocol are unavailable, so these plots do not support an aggregate CASP16 claim in the main text. The significance annotation embedded in panel (b) is retained from the original image but is not interpreted here because the paired scores and test definition are unavailable.



(a) Per-target TM-score change: RNAStructFold minus RhoFold+.



(b) TM-score distributions for the two models.

Figure 4: CASP16 TM-score comparisons, preserving the original panels. In (a), positive differences favor RNAStructFold, the opposite subtraction order to Figure 2. Panel (b) compares model-wise distributions; it is not a length-stratified analysis.