GENERALIZABLE RNA 3D STRUCTURE PREDICTION WITH RNA-ADAPTED PROTENIX AND AGENTIC TREE SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting RNA three-dimensional structure remains challenging because of limited experimental data, conformational flexibility, and reduced accuracy on long and topologically complex targets. We adapt Protenix for RNA-focused structure prediction using nucleotide-specific inputs, RNA structural supervision, and multi-conformation training.

Our approach has three main components. First, multi-conformation supervision retains up to five experimentally observed reference conformations per RNA sequence and optimizes primarily against the best-matching valid state while retaining an auxiliary reference loss. Second, we augment the curated experimental corpus with quality-filtered synthetic RNA structures generated using a diffusion-inspired structural generation pipeline. Third, we use Agentic Tree Search to select five structures from 15 candidates generated by RNA-adapted Protenix, template-based matching, and Boltz, balancing source-model reliability, conformational diversity, and structural priors.

RNA-adapted Protenix is trained with MMseqs2-derived evolutionary information, while benchmark inference is performed without inference-time MSA under the competition environment. The strongest standalone configuration reaches TM-scores of 0.489 on the public benchmark and 0.463 on the hidden private benchmark of the Stanford RNA 3D Folding Challenge. Agentic Tree Search increases performance to 0.684 public and 0.632 private. Synthetic augmentation improves the public score further but does not improve the private score, highlighting the importance of evaluating structural generalization on hidden RNA targets.

1 INTRODUCTION

RNA three-dimensional structure prediction remains difficult because of sparse experimental structures, conformational flexibility, and long-range tertiary interactions. Alternative structural states further complicate learning from single-reference supervision, while blind evaluations such as CASP15, CASP16, and RNA-Puzzles continue to expose errors in global topology and flexible RNA motifs (Baek et al., 2024b; Townshend et al., 2025; Westhof et al., 2025).

Recent biomolecular structure predictors increasingly rely on geometric learning and diffusion-based generative modeling. AlphaFold 3 extends diffusion-based structure prediction to heterogeneous biomolecular systems, including RNA (Abramson et al., 2024), while Protenix provides an open-source AlphaFold 3-style framework supporting RNA and molecular complexes (Chen et al., 2025). These developments provide a strong basis for adapting general biomolecular structure models specifically to RNA, where structural data scarcity, conformational heterogeneity, and evolutionary information remain important challenges.

We study this problem using the Stanford RNA 3D Folding benchmark (He et al., 2025). The benchmark requires five predicted structures for each RNA target and evaluates the three-dimensional coordinates of the C1′ atom of every nucleotide. This setting makes multi-conformation prediction central to the task. Predictions are aligned using USAlign (Zhang et al., 2022), and TM-score is used as the primary measure of global structural similarity.

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.

Our primary learned predictor is an RNA-adapted version of Protenix (Chen et al., 2025). We fine-tune the model on a curated RNA structural corpus using nucleotide-specific representations, MMseqs2-derived evolutionary information during training, and supervision focused on C1′ geometry. Final competition inference is performed without inference-time MSA.

A central component of the training strategy is multi-conformation supervision. RNA molecules can occupy multiple experimentally supported structural states, yet conventional training typically assigns each sequence a single reference structure. We therefore retain up to five experimentally observed conformations for sequences with structural heterogeneity. During optimization, a separate structural loss is computed for each valid reference state. The minimum-loss conformation provides the primary supervision signal, while a small auxiliary loss on the first reference conformation provides a fixed regularizing anchor. This allows the model to adaptively match alternative experimentally supported RNA states without forcing every prediction toward a single nominal reference conformation.

To increase structural diversity during training, we further augment the curated experimental corpus with quality-filtered synthetic RNA structures generated using a diffusion-based structural generation pipeline inspired by RFdiffusion (Watson et al., 2023). Synthetic examples expand the range of conformational geometries seen during fine-tuning while experimentally resolved structures remain the principal supervision source.

We additionally evaluate DRfold (Li et al., 2023) as a geometric RNA structure baseline. Because its native output does not directly provide the C1′ sugar coordinates required by the benchmark, we develop a ResidueGNN module that predicts C1′ positions from DRfold-derived RNA geometry. This provides a benchmark-compatible geometric baseline while remaining separate from the final ensemble pipeline.

At inference time, the final framework combines three complementary sources of structural hypotheses. RNA-adapted Protenix contributes candidates learned from the RNA-specific training corpus, template-based matching (TBM) contributes template-guided global topology, and Boltz (Wohlwend et al., 2024) contributes independent predictions from its publicly released pretrained weights. Boltz is used without RNA-specific retraining. Each model generates five candidate conformations, resulting in a pool of 15 structures for each target.

The final five predictions are selected using Agentic Tree Search (ATS). ATS balances model reliability, structural diversity, and consistency with fixed structural priors. Diversity is measured using RMSD after optimal Kabsch alignment (Kabsch, 1976), which removes arbitrary rigid-body rotation and translation before comparing conformations. Unlike static ranking, ATS recomputes candidate utilities after every selection because the diversity of a remaining structure depends on the conformations already selected. The procedure therefore performs adaptive, state-dependent search over the candidate pool rather than ranking all predictions independently.

The main contributions of this work are:

- an RNA-adapted Protenix pipeline trained on a curated RNA structural corpus with evolutionary information and C1′-focused supervision;
- multi-conformation learning using up to five experimentally observed reference conformations per RNA sequence;
- synthetic RNA structural augmentation and a ResidueGNN module that converts DRfold-derived geometry into benchmark-compatible C1′ coordinates; and
- Agentic Tree Search for adaptively selecting five complementary conformations from Protenix, TBM, and Boltz candidates.

2 Related Work

Recent RNA structure prediction methods span sequence-based representation learning, geometric deep learning, diffusion modeling, and end-to-end biomolecular structure prediction. Ribonanza and RibonanzaNet demonstrated the value of large-scale RNA data generation and learned sequence representations for RNA modeling (He et al., 2024). RhoFold+ further showed that pretrained RNA language models can support accurate tertiary structure prediction from sequence (Shen et al., 2024). In

parallel, geometric approaches such as ARES demonstrated that symmetry-aware learning directly from 3D coordinates can remain effective even in low-data regimes (Townshend et al., 2021).

Several frameworks have extended structure prediction beyond proteins. RoseTTAFoldNA introduced joint modeling of proteins and nucleic acids within an end-to-end architecture (Baek et al., 2024a). AlphaFold 3 introduced a diffusion-based generative architecture for heterogeneous biomolecular complexes, including RNA, proteins, ligands, and modified residues (Abramson et al., 2024). Protenix subsequently provided an open-source AlphaFold 3-style framework supporting proteins, RNA, and molecular complexes (Chen et al., 2025). Our work builds on this framework but focuses specifically on adapting its training and inference pipeline for RNA-only structure prediction.

Community-wide blind evaluations continue to highlight the difficulty of RNA tertiary structure prediction. CASP15 established RNA structure prediction as a formal assessment category and exposed persistent errors in junctions, pseudoknots, and global fold topology (Baek et al., 2024b). CASP16 further increased the difficulty through longer targets and reduced homologous information (Townshend et al., 2025). RNA-Puzzles similarly demonstrated that ligand-bound RNAs, flexible motifs, and complex tertiary arrangements remain challenging for both learning-based and traditional methods (Westhof et al., 2025).

The Stanford RNA 3D Folding Challenge provided an additional large-scale blind benchmark for automated RNA coordinate prediction (He et al., 2025). The competition used separate public and private evaluation targets, with the private set remaining hidden from participants during model development. This makes the benchmark particularly useful for studying whether improvements observed on public development targets translate to unseen and structurally diverse RNAs. We use this setting to analyze the interaction between training-data composition, conformational supervision, evolutionary information, and search-based inference.

3 METHOD

3.1 OVERVIEW

Our framework combines four stages: curated RNA structural data, synthetic structural augmentation, multi-conformation supervision, and adaptive multi-model conformer selection. Figure 1 summarizes the complete pipeline.

Experimentally resolved RNA structures are first curated to construct an RNA-specific training corpus. For sequences observed in multiple structural states, up to five experimentally determined conformations are retained. The experimental corpus is further supplemented with quality-filtered synthetic RNA structures to increase structural diversity.

We adapt Protenix (Chen et al., 2025) for RNA-focused structure prediction and train it using nucleotide-specific inputs, MSA-derived evolutionary information during training, and multi-conformation supervision. Final competition-time Protenix inference is performed without inference-time MSA. During inference, RNA-adapted Protenix, template-based matching (TBM), and Boltz (Wohlwend et al., 2024) each generate five candidate structures, producing a pool of 15 conformations for each target. Agentic Tree Search (ATS) then selects five final conformations by balancing source-model reliability, structural diversity, and consistency with structural priors.

3.2 RNA-ADAPTED PROTENIX

Protenix is an open-source biomolecular structure prediction framework based on the AlphaFold 3 modeling paradigm (Chen et al., 2025). We adapt its training pipeline specifically for RNA structure prediction and fine-tune the model on our curated RNA structural corpus.

RNA sequences are represented over the nucleotide alphabet $\{A, U, G, C\}$, and RNA-specific embeddings are optimized during training. Evolutionary information is incorporated through multiple sequence alignments generated using MMseqs2. Homologous RNA sequences are retrieved when available and supplied as evolutionary context during training.

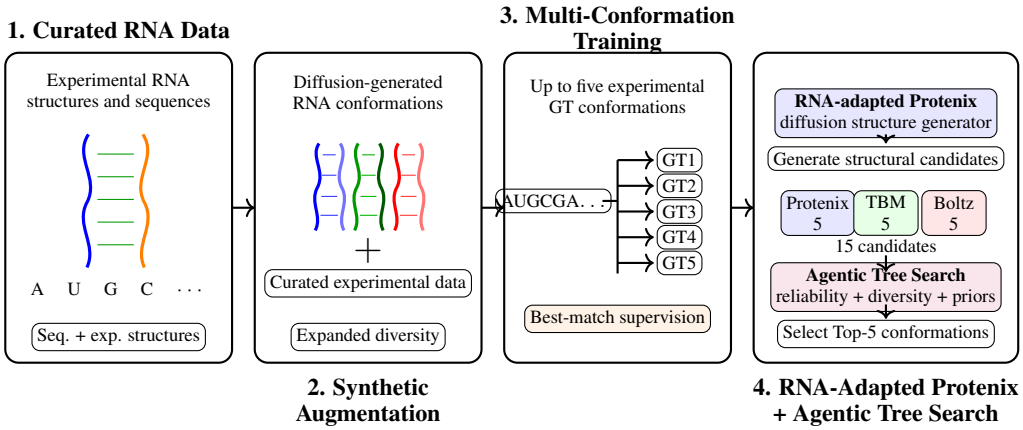


Figure 1: Overview of the proposed RNA 3D structure prediction framework. Experimental and synthetic RNA structures are used for training with up to five experimentally observed reference conformations per sequence. RNA-adapted Protenix is trained with MSA-derived evolutionary information, while final competition inference is performed without inference-time MSA. Protenix, TBM, and Boltz generate candidate structures, and Agentic Tree Search selects the final top-5 conformations.

In the competition setting, final Protenix inference is performed without inference-time MSA. Online MSA generation is unavailable in the Kaggle submission environment, and the reported benchmark predictions are therefore generated directly from the RNA sequence using the fine-tuned model. This distinguishes MSA-assisted training from MSA-free competition inference.

The benchmark evaluates C1′ coordinates for each nucleotide, and our training pipeline therefore places particular emphasis on accurate RNA backbone geometry around the C1′ representation. Rather than restricting every RNA sequence to a single structural target, the adapted training procedure supports multiple experimentally observed conformations for the same sequence.

Detailed preprocessing, MSA construction, and additional Protenix configuration settings are provided in the appendix.

3.3 Multi-Conformation Supervision

RNA molecules can occupy multiple experimentally supported conformational states. Consequently, an RNA sequence observed experimentally in more than one structural configuration should not necessarily be optimized against a single arbitrarily selected reference structure.

For each RNA sequence x, we retain up to five experimentally observed reference conformations,

$$\mathcal{G}(x) = \{G_1, G_2, \ldots, G_K\}, \qquad K \leq 5. \tag{1}$$

Given a predicted structure $\hat{G}$, the structural loss is evaluated against each available reference state,

$$\mathcal{L}_k = \mathcal{L}_{\text{struct}}(\hat{G}, G_k), \qquad k \in \{1, \ldots, K\}. \tag{2}$$

The best-matching reference state is selected dynamically as

$$k^* = \arg \min_{k \in \{1,\ldots,K\}} \mathcal{L}_k. \tag{3}$$

The implemented multi-conformation objective is

$$\mathcal{L}_{\text{multi}} = \min_{k \in \{1,\ldots,K\}} \mathcal{L}_k + 0.1\, \mathcal{L}_1. \tag{4}$$

Equivalently,

$$\mathcal{L}_{\text{multi}} = \mathcal{L}_{k^*} + 0.1\,\mathcal{L}_1. \tag{5}$$

The minimum-loss term allows the model to receive its primary supervision from the experimentally observed conformation most compatible with the current prediction. Because the identity of k^* can change during optimization, the supervision is dynamic rather than permanently tied to one structural state. The auxiliary $0.1\,\mathcal{L}_1$ term retains a small contribution from the first reference conformation, providing a fixed regularizing anchor.

This formulation reduces the penalty incurred when a prediction corresponds to an experimentally supported alternative RNA conformation rather than the nominal first reference. It does not require a single prediction to reproduce all conformations simultaneously; instead, the dominant supervision signal can follow the most compatible valid state.

3.4 Synthetic Structural Augmentation

Because experimentally resolved RNA structures remain limited, we augment the training corpus with synthetic conformations generated using a diffusion-based procedure inspired by RFdiffusion (Watson et al., 2023). RFdiffusion was originally developed for proteins; here, only the diffusion-based generation principle is adapted to construct additional RNA training examples.

From 3,982 generated candidates, structural and quality filtering retains 400 synthetic RNA sequences with associated conformations. These are used only for training and not as benchmark evaluation targets. The augmentation broadens the conformational diversity seen during training while experimental structures remain the primary source of supervision. Generation and filtering details are provided in the appendix.

3.5 Candidate Structure Generation

At inference time, we obtain structural hypotheses from three complementary prediction pipelines: RNA-adapted Protenix, template-based matching (TBM), and Boltz.

RNA-adapted Protenix generates five conformations using the RNA-specific model described above. Competition-time Protenix inference is performed without inference-time MSA.

TBM generates five template-guided structural hypotheses using a pipeline inspired by the publicly released RNA 3D Folds TBM-only notebook (Rao, 2025). Our implementation introduces additional template scoring, geometric refinement, structural variant generation, and diversity-aware candidate selection; details are provided in Appendix H.

Boltz contributes five additional predictions using its publicly released pretrained weights (Wohlwend et al., 2024). Boltz is used as an independent open-source predictor and is not retrained or fine-tuned on the custom RNA training corpus developed in this work. Boltz inference is also performed without inference-time MSA in the competition environment.

The complete candidate pool is

$$\mathcal{C} = \mathcal{C}_P \cup \mathcal{C}_T \cup \mathcal{C}_B, \qquad |\mathcal{C}| = 15, \tag{6}$$

where $\mathcal{C}_P$, $\mathcal{C}_T$, and $\mathcal{C}_B$ denote the five Protenix, TBM, and Boltz conformations, respectively.

These predictors contribute complementary structural information: RNA-adapted Protenix provides learned RNA-specific predictions, TBM contributes template-guided global topology, and Boltz provides independently generated structural hypotheses. Agentic Tree Search operates on this fixed 15-candidate pool and selects the final five conformations.

Algorithm 1 Agentic Tree Search for RNA conformer selection

Require: $\mathcal{C}_P$: five Protenix conformations; $\mathcal{C}_T$: five TBM conformations; $\mathcal{C}_B$: five Boltz conformations
Ensure: Five selected conformations $\mathcal{S}$
$\mathcal{C} \leftarrow \mathcal{C}_P \cup \mathcal{C}_T \cup \mathcal{C}_B$
Initialize structural prior set $\mathcal{A}$
$\mathcal{S} \leftarrow \mathcal{A}$
Set model reliability weights: $w_P = 0.75$, $w_T = 0.15$, $w_B = 0.10$
while $|\mathcal{S}| < 5$ **do**
  for each $c \in \mathcal{C} \setminus \mathcal{S}$ **do**
    $D(c, \mathcal{S}) \leftarrow$ mean Kabsch-RMSD to $\mathcal{S}$
    $P(c) \leftarrow$ mean Kabsch-RMSD to $\mathcal{A}$
    $\mathrm{Score}(c \mid \mathcal{S}) \leftarrow w_{\mathrm{model}}(c)\,[w_{\mathrm{div}} D(c, \mathcal{S}) - w_{\mathrm{dist}} P(c)]$
  end for
  $c^* \leftarrow \arg\max_{c \in \mathcal{C} \setminus \mathcal{S}} \mathrm{Score}(c \mid \mathcal{S})$
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{c^*\}$
end while
return $\mathcal{S}$

3.6 Agentic Tree Search

For each target, Protenix, TBM, and Boltz contribute five candidate conformations each, yielding a fixed pool of 15 structures. Agentic Tree Search (ATS) selects the final five conformations from this pool using state-dependent scoring.

The search is initialized from two fixed structural priors. Candidate selection balances source-model reliability, structural diversity relative to the already selected conformations, and geometric consistency with the priors. Structural comparison is performed using RMSD after Kabsch alignment (Kabsch, 1976).

Source-model reliability is encoded using fixed weights of 0.75 for Protenix, 0.15 for TBM, and 0.10 for Boltz. For candidate c and current selected set $\mathcal{S}$, $D(c, \mathcal{S})$ denotes the mean Kabsch-aligned RMSD between c and the selected conformations, while $P(c)$ denotes the mean Kabsch-aligned RMSD between c and the fixed structural priors. The state-dependent candidate score is

$$\mathrm{Score}(c \mid \mathcal{S}) = w_{\mathrm{model}}(c)\,[w_{\mathrm{div}} D(c, \mathcal{S}) - w_{\mathrm{dist}} P(c)]\,, \tag{7}$$

where $w_{\mathrm{model}}(c)$ is 0.75, 0.15, or 0.10 for Protenix, TBM, or Boltz candidates, respectively. The coefficients w_{div} and w_{dist} control the reward for structural diversity and the penalty for drifting from the structural priors.

After each selection, $\mathcal{S}$ changes and the scores of all remaining candidates are recomputed. A candidate that is diverse at one search state can therefore become redundant after a similar conformation has already been selected. This makes ATS state dependent rather than a fixed one-pass ranking.

Algorithm 1 summarizes the procedure.

The current implementation follows the highest-scoring expansion greedily rather than maintaining multiple simultaneous branches. Each selected subset, however, defines a new search state and induces a new ranking of the remaining candidates. Native coordinates, native TM-scores, and ground-truth RMSDs are not used during ATS selection.

4 Experimental Study

4.1 Datasets and Benchmark

Training data are constructed from experimentally resolved RNA structures, Stanford RNA competition resources, a PDB-derived multi-conformation dataset, and quality-filtered synthetic RNA structures. The PDB-derived corpus is created through a multi-stage preprocessing pipeline in which

RNA chains are extracted from mmCIF files, incomplete or unsupported structures are removed, and structurally distinct conformations are identified using TM-score.

The initial PDB collection contains 8,563 entries with RNA, corresponding to 5,964 unique RNA sequences. After filtering unsupported modified residues and incomplete structures, 3,203 unique sequences remain. Among these, 3,009 sequences exhibit multiple experimentally observed conformations. Subsequent DLS preprocessing and quality filtering retain 1,462 sequences for training, with up to five valid conformations retained per sequence for multi-conformation supervision.

Table 1 summarizes the principal source datasets used during model development. Counts for StanfordRNA-v1, StanfordRNA-v2, the PDB-derived DLS set, and synthetic augmentation are reported before cross-dataset merging. These source datasets are not disjoint: some DLS sequences are already present in StanfordRNA-v1 or StanfordRNA-v2, and overlap may also occur between other sources. Consequently, the source counts should not be summed directly. Duplicate RNA sequences are removed when constructing the merged training pools.

Table 1: Principal RNA datasets used during model development. Source-dataset counts are reported before cross-dataset merging and are therefore not additive. Overlapping RNA sequences, including DLS sequences already present in the Stanford RNA resources, are deduplicated when constructing Merged-v1 and Merged-v2. The merged counts denote the final deduplicated training configurations.

Dataset	Sequence / candidate count	Description
StanfordRNA-v1	784	Source RNA sequences
StanfordRNA-v2	4,214	Source RNA sequences
PDB-derived DLS set	1,462	Sequences retained after preprocessing
Synthetic augmentation	400	Retained synthetic RNA sequences with conformations
Merged-v1	5,715	Deduplicated training configuration
Merged-v2	6,090	Deduplicated training configuration

The synthetic dataset is obtained from 3,982 generated candidates, of which 400 RNA sequences with associated conformations pass structural and quality filtering and are retained for training. Merged-v1 denotes the initial combined training configuration. It includes approximately 40 publicly released CASP16-derived RNA sequences and does not include RFdiffusion-inspired synthetic augmentation. Merged-v2 denotes the revised training configuration: the CASP16-derived samples are removed and the 400 quality-filtered synthetic RNA sequences are incorporated. Cross-dataset duplicates are removed in both merged configurations.

CASP15, RNA-Puzzles, and YangLab30 targets are excluded from structural training using their known PDB identifiers and are reserved for external evaluation. The structural corpus is restricted to PDB entries released up to March 26, 2025. We do not additionally claim a global sequence-identity decontamination threshold between all training and evaluation sequences; therefore, homologous sequences may remain even though the benchmark structures themselves are excluded.

Evaluation on the Stanford RNA 3D Folding benchmark follows the competition protocol (He et al., 2025). Each target requires five predicted RNA structures represented by three-dimensional C1′ coordinates. Predictions are aligned using USAlign (Zhang et al., 2022), and TM-score is used as the primary measure of global structural similarity.

The complete filtering and conformational grouping procedures are provided in Appendix A, with leakage-control details in Appendix A.2. Additional transfer results on CASP15, RNA-Puzzles, and YangLab30 are reported in Appendix K.

4.2 Experimental Settings

RNA-adapted Protenix is fine-tuned on four NVIDIA H100 GPUs, with one RNA example per GPU and an effective batch size of four. We use Adam with learning rate 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 10^{-8}, 50 warm-up steps, gradient clipping at 10, EMA decay 0.995, and bfloat16 precision. Training runs for 4,000 optimizer steps with a maximum sequence length of 416 nucleotides.

Table 2: Main results on the Stanford RNA 3D Folding benchmark. Dynamic denotes multi-conformation supervision. Train MSA and Infer. MSA distinguish evolutionary information used during training and prediction, respectively. Protenix is trained with MSA where indicated but evaluated without inference-time MSA because online MSA generation was unavailable in the Kaggle environment. Boltz is also evaluated without inference-time MSA. Public and private columns report TM-score.

ID	Method	Training / selection	Dynamic	Train MSA	Infer. MSA	Max len.	Public	Private
P1	RNA-adapted Protenix	Merged-v1	×	×	×	800	0.420	0.381
P2	RNA-adapted Protenix	Merged-v1	✓	✓	×	800	0.448	0.424
P3	RNA-adapted Protenix	Merged-v2	✓	✓	×	800	0.489	0.463
A1	AlphaFold	Organizer reference	–	✓	✓	Full sequence	0.349	0.464
A2	AlphaFold	Organizer reference	–	×	×	Full sequence	0.299	0.376
DR1	DRfold	Open pretrained weights	–	–	–	480	0.277	0.230
DR2	DRfold + ResidueGNN	Open weights + C1′ reconstruction	–	–	–	480	0.355	0.354
B1	Boltz-v1	Open pretrained weights	–	–	×	800	0.350	0.417
E1	Protenix + Boltz + DRfold	Direct ensemble	–	Mixed	×	800†	0.487	0.483
E2	TBM + Protenix + Boltz + DRfold	Direct ensemble	–	Mixed	×	800†	0.594	0.577
S1	Protenix + TBM + Boltz	Kabsch-based top-confidence selection	–	Mixed	×	800†	0.601	0.591
ATS1	Protenix + TBM + Boltz	15 candidates (5+5+5), ATS selects 5	–	Mixed	×	800†	0.684	0.632
ATS2	Protenix + TBM + Boltz	15 candidates (5+5+5), ATS selects 5	–	Mixed	×	800†	0.692	0.620

†Sequence-length handling follows the contributing models: DRfold is limited to 480 nucleotides, while Protenix and Boltz are evaluated up to 800 nucleotides. DRfold is not part of the S1 or ATS candidate pools.

At inference, Protenix and Boltz are evaluated up to 800 nucleotides and DRfold up to 480 nucleotides, while the organizer-provided AlphaFold reference uses the full sequence. MSA-enabled Protenix models use MMseqs2-derived evolutionary information during training but not during competition inference. Kaggle notebooks were limited to an 8-hour runtime and had no internet access, preventing online MSA generation from the external sequence databases. Reported Protenix scores therefore correspond to MSA-assisted training with MSA-free deployment.

Protenix, TBM, and Boltz each generate five candidate conformations, which are processed by Agentic Tree Search as described in Section 3.6. MSA construction is detailed in Appendix B, while complete optimizer, architecture, diffusion, loss, and inference settings are provided in Appendix E.

4.3 Main Results

Table 2 summarizes the principal standalone and ensemble results on the Stanford RNA 3D Folding benchmark. Experiment identifiers use P for RNA-adapted Protenix, A for the organizer-provided AlphaFold reference, DR for DRfold-based configurations, B for Boltz, E for direct ensembles, and ATS for Agentic Tree Search.

The first three experiments summarize RNA-adapted Protenix. P1 reaches 0.420/0.381 without multi-conformation supervision or MSA. P2 adds dynamic multi-conformation supervision and MSA-assisted training, improving to 0.448/0.424. P3 uses Merged-v2 and reaches the strongest standalone Protenix result of 0.489/0.463. Merged-v1 includes approximately 40 CASP16-derived RNA sequences without synthetic augmentation, whereas Merged-v2 removes these samples and includes quality-filtered RFdiffusion-inspired synthetic structures. Matched MSA and multi-conformation ablations are provided in the appendix.

The organizer-provided AlphaFold reference reaches 0.349/0.464 with MSA and 0.299/0.376 without MSA. It uses the full input sequence and is therefore not strictly compute-matched to the competition-time inference limits of our models. DRfold reaches 0.277/0.230, while adding ResidueGNN-based C1′ reconstruction increases performance to 0.355/0.354. Boltz-v1, evaluated from publicly released pretrained weights without inference-time MSA, reaches 0.350/0.417.

Combining complementary predictors further improves performance. The direct Protenix–Boltz–DRfold ensemble reaches 0.487/0.483, while adding TBM increases the result to 0.594/0.577.

On the same Protenix–TBM–Boltz candidate pool, the simpler Kabsch-based selector reaches 0.601/0.591, while ATS improves this to 0.684/0.632.

For ATS, each target receives 15 candidates: five each from Protenix, TBM, and Boltz. ATS selects the final five, with DRfold excluded from this pool. The reference configuration reaches 0.684 public and 0.632 private, while an alternative reaches 0.692/0.620. Because only aggregate private

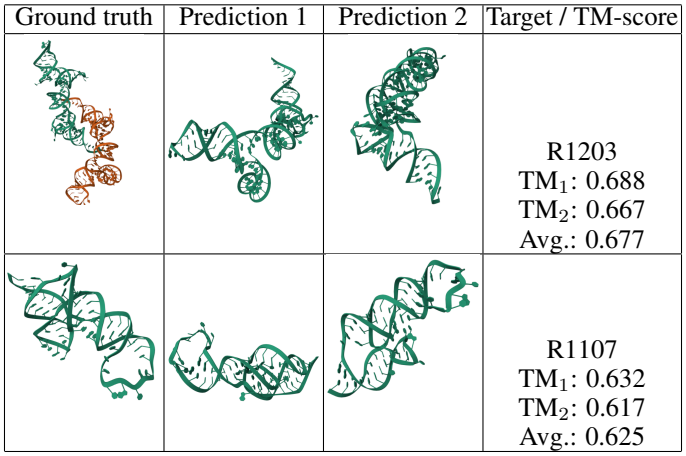


Figure 2: Representative RNA structure predictions from CASP16 (R1203) and CASP15 (R1107). Predicted conformations are aligned to the experimental reference using USAlign for evaluation and visualization. TM-score measures global structural agreement.

scores are available, small differences between these configurations are not interpreted as statistically significant.

Overall, private performance increases from 0.463 for standalone Protenix to 0.577 for the TBM-assisted ensemble and 0.632 with ATS. A post-training evaluation on 13 RNA structures released after the March 26, 2025 cutoff further shows complementary behavior between TBM and Protenix (Appendix J).

4.4 Qualitative Structural Analysis

Figure 2 shows representative predictions from CASP16 and CASP15. For CASP16 target R1203, the two displayed conformations reach TM-scores of 0.688 and 0.667, with an average of 0.677. For CASP15 target R1107, the corresponding scores are 0.632 and 0.617, with an average of 0.625. Additional examples are provided in Appendix L.

5 Discussion and Conclusion

Our results show that RNA-specific training and adaptive candidate selection provide complementary gains. Matched ablations demonstrate that both multi-conformation supervision and MSA-assisted training improve Protenix, while final competition inference remains MSA-free. The strongest standalone Protenix configuration reaches 0.489 public and 0.463 private TM-score, and combining complementary structural hypotheses further improves performance.

Agentic Tree Search operates on a fixed pool of 15 candidates from Protenix, TBM, and Boltz and selects five conformations using model reliability, structural diversity, and structural priors. The resulting configuration reaches 0.684 public and 0.632 private, compared with 0.594/0.577 for the TBM-assisted direct ensemble. Sensitivity experiments further show that ATS performance depends on source weights and anchor selection, indicating that these choices remain important components of the current search procedure.

Synthetic augmentation improves the public score but does not improve the corresponding private score, and the hidden private benchmark does not provide per-target scores for statistical testing. We therefore avoid interpreting small private-score differences as significant. Overall, the results suggest that multi-state supervision, complementary structural predictors, and state-dependent conformer selection are promising directions for improving RNA 3D structure prediction on unseen targets.

Reproducibility

Training hyperparameters, preprocessing, and dataset details are specified in Appendices A–E; code and model checkpoints will be released upon publication.

AI Use Statement

Generative AI tools were used only to aid and polish the writing; all scientific content, analyses, and conclusions were reviewed and verified by the authors.

References

Josh Abramson, Jonas Adler, Jack Dunger, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8012):493–500, 2024. doi: 10.1038/s41586-024-07487-w.

Minkyung Baek, Ryan McHugh, Ivan Anishchenko, et al. Accurate prediction of protein–nucleic acid complexes using rosettafoldna. *Nature Methods*, 21(1):117–121, 2024a. doi: 10.1038/s41592-023-02086-5.

Minkyung Baek et al. An rna structure category in casp15 reveals new capabilities and bottlenecks in rna 3d structure prediction. *bioRxiv*, 2024b. doi: 10.1101/2024.05.29.656942.

Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, Shenghao Wu, Kuangqi Zhou, Yanping Yang, Zhenyu Liu, Lan Wang, Bo Shi, Shaochen Shi, and Wenzhi Xiao. Protenix: Advancing structure prediction through a comprehensive alphafold3 reproduction. *bioRxiv*, 2025. doi: 10.1101/2025.01.08.631967.

José A. Cruz, Marie-France Blanchet, Michal Boniecki, Janusz M. Bujnicki, Shi-Jie Chen, Song Cao, Rhiju Das, Feng Ding, Nikolay V. Dokholyan, Samuel C. Flores, Liang Huang, Christopher A. Lavender, Valentina Lisi, François Major, Krzysztof Mikolajczak, Dilip J. Patel, Andrew Philips, Tomasz Puton, John Santalucia, Fabien Sijenyi, Thomas Hermann, Kristian Rother, Michal Rother, Alexander Serganov, Michal Skorupski, Tomasz Soltysinski, Parin Sripakdeevong, Izabela Tuszynska, Kevin M. Weeks, Christina Waldsich, Marion 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.

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.

Shujun He, Rui Huang, Jill Townley, et al. Ribonanza: Deep learning of rna structure through dual crowdsourcing. *bioRxiv*, 2024. doi: 10.1101/2024.02.24.581671.

Shujun He, CASP16 Organizers, CASP16 RNA Experimentalists, RNA-Puzzles Consortium, VFOLD Team, Rachael Kretsch, Alissa Hummer, Andrew Favor, Walter Reade, Maggie Demkin, Rhiju Das, et al. Stanford rna 3d folding. Kaggle Competition, 2025. URL `https://www.kaggle.com/competitions/stanford-rna-3d-folding`.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976. doi: 10.1107/S0567739476001873.

Yang Li, Chengxin Zhang, Chenjie Feng, Robin Pearce, P. Lydia Freddolino, and Yang Zhang. Integrating end-to-end learning with deep geometrical potentials for ab initio rna structure prediction. *Nature Communications*, 14(1):5745, 2023. doi: 10.1038/s41467-023-41385-6.

G. John Rao. Rna 3d folds: Tbm-only approach. Kaggle Notebook, 2025. URL `https://www.kaggle.com/code/jaejohn/rna-3d-folds-tbm-only-approach`. Stanford RNA 3D Folding Challenge.

Tao Shen, Zhihang Hu, Siqi Sun, et al. 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.

Raphael J. L. Townshend, Stephan Eismann, Andrew M. Watkins, et al. Geometric deep learning of rna structure. *Science*, 373(6558):1047–1051, 2021. doi: 10.1126/science.abe5650.

Raphael J. L. Townshend et al. Rna structure prediction in casp16 reveals new methods, data limitations and emerging directions. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707.

Wenkai Wang, Chenjie Feng, Rui Han, et al. trrosettarna: automated prediction of rna 3d structure with transformer network. *Nature Communications*, 14:7266, 2023. doi: 10.1038/s41467-023-42528-4.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with rfdiffusion. *Nature*, 620:1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.

Eric Westhof, Zhichao Miao, et al. Rna-puzzles round v: Blind predictions of 23 rna structures. *Nature Methods*, 22:399–411, 2025. doi: 10.1038/s41592-024-02543-9.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, 2024. doi: 10.1101/2024.11.19.624167.

Chengxin Zhang, Morgan Shine, Anna Marie Pyle, and Yang Zhang. Us-align: Universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature Methods*, 19(9): 1109–1115, 2022. doi: 10.1038/s41592-022-01585-1.

A Dataset Curation Details

The main paper summarizes the datasets used during model development. Here we describe the underlying RNA structure filtering, conformational grouping, and leakage-control procedure.

A.1 RNA Structure Filtering

The RNA structural corpus is constructed from experimentally resolved structures available in the RCSB PDB up to March 26, 2025. The initial collection contains 8,563 PDB entries corresponding to 23,630 RNA chains and 5,964 unique sequences. Chains containing unsupported modified residues or incomplete structural information are removed before training, while alternative experimentally observed conformations are preserved.

Figure 3 summarizes the complete curation procedure.

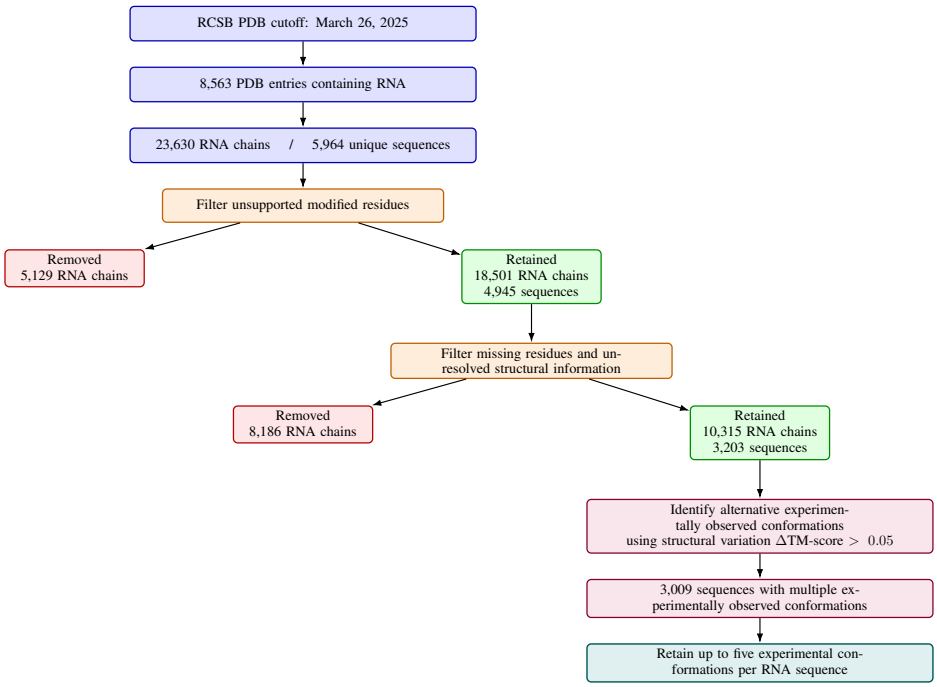


Figure 3: RNA structure curation and multi-conformation selection pipeline. Blue blocks denote the initial structural collection, orange blocks denote filtering operations, red and green blocks denote removed and retained structures, purple blocks denote conformational grouping, and the teal block denotes the final multi-conformation selection stage.

After modified-residue filtering, 18,501 RNA chains representing 4,945 unique sequences remain. Removing incomplete structures results in 10,315 RNA chains and 3,203 unique sequences. Among the curated sequences, 3,009 exhibit multiple experimentally observed structural states. Up to five conformations are retained per sequence for multi-conformation supervision.

A.2 Leakage Control

The structural training corpus is restricted to PDB entries released up to March 26, 2025. Merged-v1 includes approximately 40 publicly released CASP16 RNA sequences used during early development, whereas Merged-v2 removes these CASP16-derived samples and incorporates the quality-filtered synthetic RNA structures.

CASP15, RNA-Puzzles, and YangLab30 evaluation targets were excluded from the structural training corpus using their known PDB identifiers. These structures were reserved exclusively for external evaluation.

We did not additionally apply a global sequence-identity decontamination threshold between all training and evaluation sequences; therefore, homologous training sequences may still remain even though the benchmark structures themselves were removed.

B Evolutionary Information and MSA Construction

Evolutionary sequence resources are assembled from RNAcentral, Rfam, and the NCBI nucleotide database. These resources are used as searchable sequence databases for MMseqs2-based homolog retrieval and MSA construction. They provide evolutionary sequence context and are not used as structural supervision labels.

For each query RNA, homologous sequences are retrieved using MMseqs2 and ranked by sequence-search relevance. Up to ten non-redundant homologs are retained when available and transformed into the MSA representation supplied to RNA-adapted Protenix.

The MSA-enabled Protenix configurations use evolutionary information during training. Final competition-time inference is performed without inference-time MSA because online MSA generation is unavailable in the Kaggle submission environment. MSA-disabled experiments use the same underlying model without these evolutionary features.

The model MSA module contains four blocks with representation dimension 64, MSA dropout 0.15, and pair dropout 0.25.

C Synthetic Structural Augmentation

To increase structural diversity beyond experimentally resolved RNA structures, additional conformations are generated using a diffusion-inspired structural generation procedure. The approach is inspired by the generative principle of RFdiffusion (Watson et al., 2023); RFdiffusion itself was originally developed for proteins.

A total of 3,982 synthetic structural candidates are generated. Structural and quality filtering retain 400 RNA sequences with associated conformations for inclusion in the expanded RNA training corpus.

In the final ATS comparison, synthetic augmentation increases the public TM-score from 0.684 to 0.692, whereas the corresponding private score changes from 0.632 to 0.620. We therefore interpret augmentation as an additional source of conformational diversity rather than as a uniformly beneficial source of generalization.

D RNA-Adapted Protenix Details

RNA-adapted Protenix is initialized from the open-source Protenix framework (Chen et al., 2025) and fine-tuned on the curated RNA structural corpus.

RNA sequences are represented using nucleotide-specific inputs. MMseqs2-derived evolutionary information is incorporated in the MSA-enabled configurations.

For sequences associated with multiple experimentally observed structural states, multi-conformation supervision is used. Given a prediction $\hat{\mathbf{Y}}$ and valid reference structures $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_K\}$, the implementation selects the minimum candidate loss while retaining a small contribution from the first reference state:

$$\mathcal{L}_{\text{multi}} = \min_{k \in \{1,\ldots,K\}} \mathcal{L}(\hat{\mathbf{Y}}, \mathbf{Y}_k) + 0.1\, \mathcal{L}(\hat{\mathbf{Y}}, \mathbf{Y}_1). \tag{8}$$

This formulation allows alternative experimentally supported conformations to contribute to optimization without forcing each sequence toward a single reference geometry.

E Training and Inference Hyperparameters

Training is performed using four NVIDIA H100 GPUs. Each GPU processes one independent RNA example per optimization step, giving an effective RNA example batch size of four. The configured diffusion batch size controls noisy diffusion replicas rather than the number of independent RNA molecules.

Table 3: Training and optimization settings for RNA-adapted Protenix. Training is distributed across four NVIDIA H100 GPUs with one RNA example per GPU. The diffusion replica batch size represents multiple noisy realizations of a training structure rather than independent RNA examples.

Hyperparameter	Value
Training hardware	$4 \times$ NVIDIA H100
RNA examples per GPU	1
Effective RNA batch size	4
Gradient accumulation	1
Numerical precision	bfloat16
Optimizer	Adam
Learning rate	1×10^{-4}
Adam β_1, β_2	$(0.9, 0.95)$
Weight decay	1×10^{-8}
Gradient clipping norm	10
Learning-rate scheduler	AF3 schedule
Warm-up steps	50
Optimizer steps	4,000
EMA decay	0.995
Random seed	42
Training sequence limit	416 nt
Diffusion replica batch size	8
Checkpoint interval	2,000 steps
Logging interval	1 step

F Sequence-Length Handling

The evaluated predictors operate under different sequence-length conditions because of competition-time computational constraints.

The organizer-provided AlphaFold reference results use the complete input sequence. In contrast, Protenix and Boltz are evaluated up to 800 nucleotides under the available Kaggle inference resources, while DRfold-based predictions are limited to 480 nucleotides.

For multi-model configurations, each predictor contributes candidates according to its corresponding supported length. TBM additionally provides template-derived structural hypotheses.

G DRfold and ResidueGNN

DRfold (Li et al., 2023) is evaluated as an RNA geometric structure baseline. Its native output does not directly provide the C1′ coordinate representation required by the Stanford RNA benchmark. ResidueGNN is therefore used to map DRfold-derived geometry to the required C1′ coordinates.

Figure 4 summarizes the pipeline.

ResidueGNN is used only to construct a benchmark-compatible DRfold baseline and is not included in the final Protenix–TBM–Boltz candidate pool used by ATS.

H Template-Based Matching

Our template-based modeling pipeline is inspired by the publicly released RNA 3D Folds TBM-only notebook (Rao, 2025), but includes several implementation-specific extensions.

Table 4: Principal architecture, diffusion, loss, and inference settings. The model retains the Protenix Pairformer and diffusion architecture. Training uses 20-step diffusion sampling, while the batch-inference path uses 10 model cycles and 200 diffusion steps to generate five candidate structures per RNA target.

Parameter	Value	Parameter	Value
Single representation c_s	384	Pair representation c_z	128
Input representation	449	Token representation	384
Atom representation	128	Atom-pair representation	16
Pairformer blocks	48	Pairformer heads	16
Pairformer dropout	0.25	MSA blocks	4
MSA representation	64	MSA dropout	0.15
MSA pair dropout	0.25	Diffusion token dimension	768
Diffusion transformer blocks	24	Diffusion transformer heads	16
Atom encoder blocks / heads	3 / 4	Atom decoder blocks / heads	3 / 4
p_{mean}	-1.2	p_{std}	1.5
σ_{data}	16.0	s_{max}	160.0
s_{min}	0.0004	Noise-schedule ρ	7
γ_0	0.8	γ_{min}	1.0
Noise-scale λ	1.003	Step-scale η	1.5
Training diffusion steps	20	Mini-rollout steps	20
Mini-rollout samples	1	Confidence loss weight	1×10^{-4}
Diffusion loss weight	4.0	Distogram loss weight	0.03
Smooth lDDT weight	1.0	RNA MSE coefficient	5.0
Nucleotide lDDT neighborhood	30 Å	lDDT thresholds	0.5, 1, 2, 4 Å
Inference model cycles	10	Inference diffusion steps	200
Generated structures per target	5	Inference seed	101

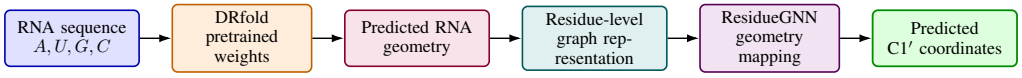


Figure 4: DRfold–ResidueGNN baseline pipeline. DRfold predicts RNA structural geometry from the input sequence using pretrained weights. The resulting residue-level geometry is converted into a graph representation and processed by ResidueGNN to generate the C1′ coordinates required by the benchmark.

For each target, candidate RNA templates are identified from the available structural training set using global sequence alignment. Templates are ranked with a composite score that combines sequence identity, query coverage, template coverage, sequence-length agreement, normalized alignment score, and a gap penalty. Up to 20 templates are retained for candidate generation.

Matched template coordinates are transferred to the query sequence. Unmatched internal regions are reconstructed by interpolation between flanking C1′ coordinates, while terminal gaps are extrapolated along the local backbone direction. The resulting structures are refined using local RNA geometry constraints, including nearest- and second-neighbor C1′–C1′ distances, smoothing, and steric-clash correction.

Multiple candidate structures are generated from high-ranking templates and their available conformations. Additional candidates are produced using controlled coordinate perturbations and mild hinge transformations followed by geometric refinement. Candidate generation is capped at 30 structures per target.

The final five TBM conformations are selected using a combination of template-derived quality and structural diversity. Diversity is measured using Kabsch-aligned RMSD relative to previously selected structures, with the diversity contribution bounded to avoid favoring unrealistic candidates solely because they are highly dissimilar.

These five TBM conformations are combined with five Protenix and five Boltz conformations to form the 15-candidate pool used by Agentic Tree Search.

I BOLTZ INFERENCE

Boltz is evaluated using its publicly released pretrained weights (Wohlwend et al., 2024). No RNA-specific retraining or fine-tuning is performed using the curated corpus developed in this work.

Five candidate conformations are generated for each supported target. Under the Kaggle inference environment, Boltz is evaluated with a maximum sequence length of 800 nucleotides.

J POST-TRAINING PDB BENCHMARK

To evaluate generalization beyond the structures available during model development, we perform a post-hoc analysis on 13 RNA structures released in the PDB after the March 26, 2025 training-data cutoff. These targets were therefore unavailable during model training.

For each target, predictions were generated independently using TBM and the RNA-adapted Protenix model. After the corresponding experimental structures became available, the generated prediction files were evaluated using TM-score. The "Best source" column reports which of the two prediction sources achieved the higher TM-score for retrospective analysis only; native structures and TM-scores were not available during prediction and were not used by Agentic Tree Search.

Table 5: Post-hoc evaluation of independently generated TBM and Protenix predictions on 13 RNA structures released in the PDB after the March 26, 2025 training cutoff. $\mathrm{TM}_{\mathrm{TBM}}$ and $\mathrm{TM}_{\mathrm{Protenix}}$ are computed only after the experimental structures became available. "Best source" indicates the higher-scoring predictor for retrospective analysis and is not used during inference.

Target ID	$\mathrm{TM}_{\mathrm{TBM}}$	$\mathrm{TM}_{\mathrm{Protenix}}$	Best source	Best TM
8VQV	0.91708	0.74352	TBM	0.91708
9E9Q	0.56803	0.60023	Protenix	0.60023
9G4P	0.46631	0.31505	TBM	0.46631
9G4Q	0.33465	0.43069	Protenix	0.43069
9G4R	0.23963	0.43340	Protenix	0.43340
9IWF	0.66022	0.59655	TBM	0.66022
9J09	0.26842	0.38660	Protenix	0.38660
9J4N	0.73014	0.74350	Protenix	0.74350
9J4O	0.69472	0.64824	TBM	0.69472
9JGM	0.66172	0.76514	Protenix	0.76514
9LKU	0.83085	0.86004	Protenix	0.86004
9MME	0.74523	0.20920	TBM	0.74523
9MMG	0.91277	0.24754	TBM	0.91277

The two predictors exhibit complementary behavior across these unseen post-training targets. Protenix performs better on seven of the 13 RNAs, whereas TBM performs better on six. In particular, TBM shows a large advantage on the two 580-nucleotide targets, 9MME and 9MMG, while Protenix performs better on several short and medium-length targets. These differences indicate that sequence length alone does not determine which prediction strategy is most effective.

This retrospective evaluation supports the use of heterogeneous structural hypotheses in the final ensemble. TBM can provide strong template-derived global topology for targets with suitable structural analogs, whereas RNA-adapted Protenix can perform better when template transfer is less effective. Their complementary behavior motivates combining both sources, together with Boltz, in the 15-candidate pool used by Agentic Tree Search.

K EXTERNAL BENCHMARK EVALUATION

The framework is additionally evaluated on YangLab30, CASP15, and RNA-Puzzles to assess transfer beyond the Stanford RNA benchmark.

Table 6: Performance on external RNA structure prediction benchmarks. Values report mean TM-score over the corresponding target collection. Because the benchmarks differ in target composition and difficulty, they are treated as separate transfer evaluations rather than directly comparable leaderboards.

Benchmark	Number of targets	TM-score
YangLab30 (Wang et al., 2023)	30	0.649
CASP15 (Das et al., 2023)	15	0.473
RNA-Puzzles (Cruz et al., 2012)	20	0.451

L Additional Qualitative Structural Results

L.1 Additional CASP16 Examples

Target R1203 is shown in the main paper. Figure 5 presents two additional CASP16 examples. Predicted conformations are aligned to their experimental references using USAlign for evaluation and visualization.

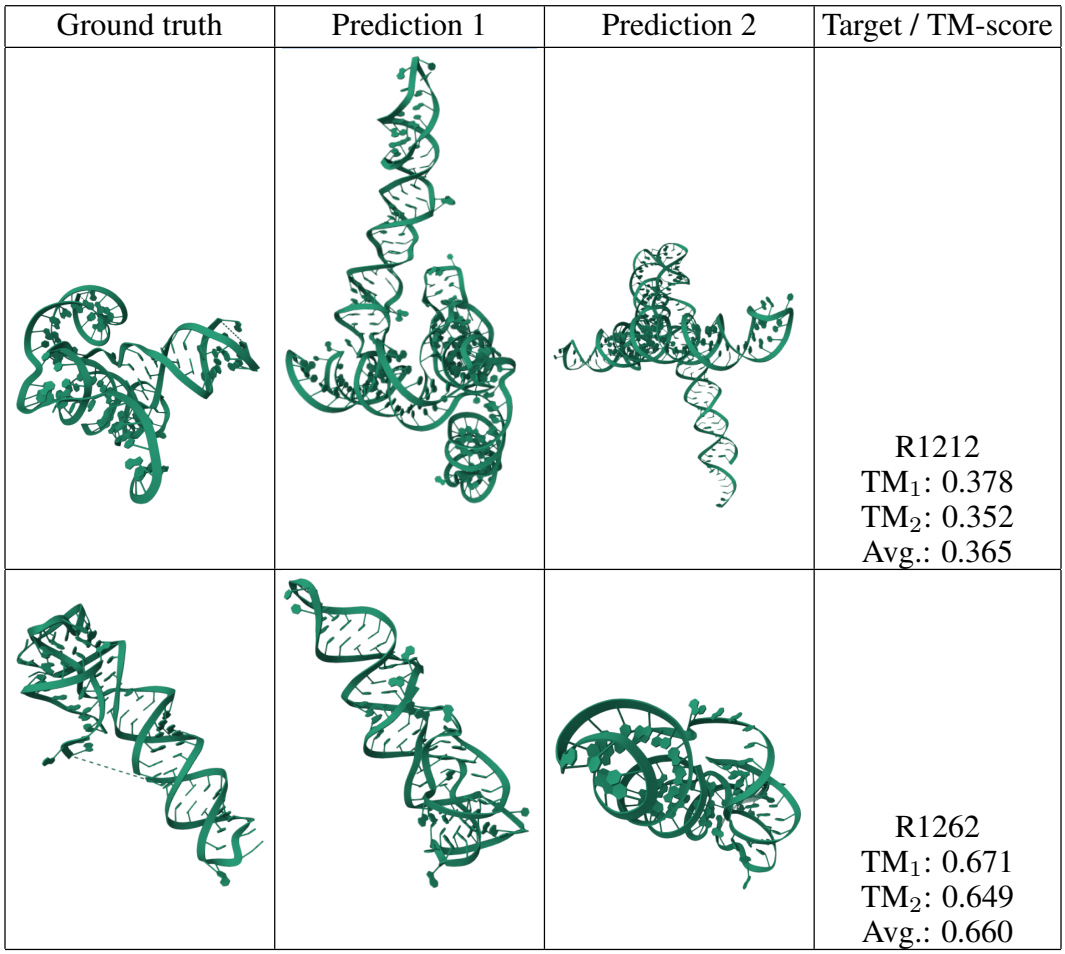


Figure 5: Additional CASP16 qualitative examples. Each row compares the experimental reference with two predicted conformations and their corresponding TM-scores.

L.2 Additional CASP15 Examples

Target R1107 is shown in the main paper. Figure 6 presents two additional CASP15 examples.

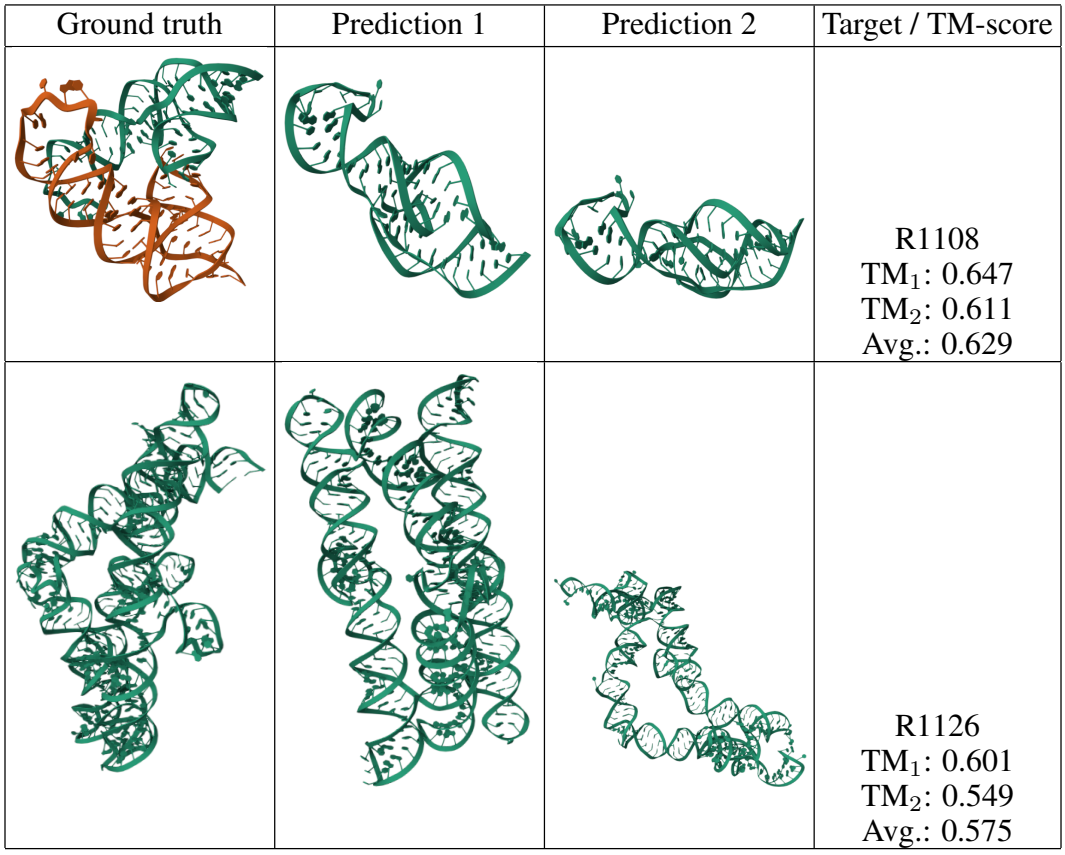


Figure 6: Additional CASP15 qualitative examples. Each row compares the experimental reference with two predicted conformations and their corresponding TM-scores.

M Additional Implementation Details

The complete computational workflow includes RNA structure preprocessing, multi-conformation grouping, MMseqs2 homolog retrieval, RNA-adapted Protenix fine-tuning, synthetic structure generation and filtering, DRfold-to-C1′ reconstruction, TBM candidate generation, Boltz inference, Kabsch-aligned candidate comparison, and Agentic Tree Search conformer selection.

USAlign (Zhang et al., 2022) is used for benchmark structural alignment and TM-score evaluation. Kabsch alignment (Kabsch, 1976) is used separately inside ATS to compute rigid-body-invariant RMSD between candidate structures.

N Additional Ablation Studies

N.1 Multi-Conformation and MSA Ablation

Table 7: Matched ablation of multi-conformation supervision and MSA-assisted training under the same underlying dataset and model configuration. All benchmark predictions are generated without inference-time MSA.

Configuration	Multi-conf.	Train MSA	Public	Private
Baseline	×	×	0.420	0.381
Multi-conf. only	✓	×	0.431	0.412
MSA only	×	✓	0.424	0.396
Multi-conf. + MSA	✓	✓	0.448	0.424

Starting from 0.420/0.381 without either component, multi-conformation supervision alone improves performance to 0.431/0.412, while MSA-assisted training alone reaches 0.424/0.396. Combining both components yields 0.448/0.424. The matched comparison therefore shows that both components contribute, with multi-conformation supervision producing the larger individual improvement on the private benchmark.

N.2 ATS Sensitivity

Table 8: Sensitivity of ATS to source-model weights and structural anchor selection.

Configuration	Public	Private
Source-model weights		
$0.75/0.15/0.10$	0.684	0.632
$0.60/0.25/0.15$	0.663	0.615
$0.50/0.25/0.25$	0.650	0.600
$1/3,\ 1/3,\ 1/3$	0.676	0.614
Anchor selection		
Protenix #2 + TBM #1	0.684	0.632
Protenix #1 + TBM #1	0.653	0.612
Protenix #2 + TBM #2	0.662	0.619

Alternative source-model weights and anchor choices produce lower aggregate scores than the reference ATS configuration. The results also show that ATS is sensitive to these design choices, motivating the use of the reported reference parameters rather than suggesting that the method is invariant to weight or anchor selection.

O Limitations

Experimentally resolved RNA structures remain substantially less abundant than protein structures, limiting the diversity of direct structural supervision. Multi-conformation training partially addresses RNA structural heterogeneity but remains constrained by the number of experimentally observed alternative states.

Synthetic augmentation broadens the conformational distribution encountered during training, but its benefit is not uniform across evaluation sets. In the reported ATS comparison, augmentation increases the public score while not increasing the corresponding private score.

Diffusion-based inference is computationally demanding for long RNA sequences. The 800-nucleotide limits used for Protenix and Boltz and the 480-nucleotide limit used for DRfold reflect competition-time computational constraints rather than intrinsic biological sequence-length limits.

Finally, the current ATS implementation performs greedy state-dependent selection rather than maintaining multiple simultaneous search branches. Candidate utilities are recomputed after each selected conformation, but the quality of the final ensemble remains dependent on the diversity and accuracy of the original candidate pool.