Oracle Selection in Reinforcement Learning-Based RNA Inverse Design

Amanda New[1] and Herbert H. Tsang[1]

[1]Applied Research Lab, Trinity Western University, Langley, British Columbia, Canada.

Contributing authors: amandanew202@gmail.com; htsang@twu.ca;

Abstract

Background: RNA inverse design seeks sequences that fold into a target three-dimensional structure. In reinforcement learning (RL)-based approaches, the folding oracle converts generated sequences into predicted structures and thereby defines the reward signal used to update the sequence-generation policy. Because independent benchmarks provide mixed evidence regarding the relative RNA structure-prediction performance of general and RNA-specialist models, it is unclear whether predictive performance on conventional benchmarks translates directly into utility as an RL reward oracle. We investigate this question in RIDER by replacing its RNA-specialist oracle with AlphaFold3 (AF3) across the 14-structure Das benchmark. **Results:** Across the benchmark, oracle choice affected both the structural self-consistency reached during RL fine-tuning and the resulting training dynamics, with RhoFold generally showing stronger performance than AF3 under the conditions tested. AF3 achieved lower peak TM-scores than RhoFold on 10 of 14 targets, with mean peak TM-scores of 0.332 and 0.437, respectively. AF3 achieved higher peak TM-scores on three short structural RNAs, while one target (1XPE) was effectively tied. An auxiliary AF3 MSA-sensitivity analysis showed that the RL performance deficit persisted on several targets for which MSA-free inference produced little change relative to MSA-enabled AF3 predictions, suggesting that MSA availability alone does not account for the observed RL differences. A reward-distribution comparison across all 14 targets showed weak agreement between oracle rewards assigned to identical candidate sequences in both magnitude (mean Pearson $\boldsymbol{r = 0.103}$, median $\mathbf{0.091}$) and ranking (mean Spearman $\boldsymbol{\rho = 0.069}$, median $\mathbf{0.085}$). Three targets (1L2X, 1Q9A, 1LNT) survived Benjamini–Hochberg correction for both statistics; notably, 1Q9A showed strong inverse rank agreement ($\boldsymbol{\rho = -0.738}$). Reward dispersion was highly target-dependent: median AF3/RhoFold variance, MAD, and IQR ratios were 0.98, 0.80, and 0.86, respectively, despite means above 1 because of a small number of high-dispersion targets. The strongest exploratory association with RL outcome was observed for the log MAD ratio ($\boldsymbol{r = -0.622}$, $\boldsymbol{p = 0.017}$), but this did not survive correction across the three dispersion measures ($\boldsymbol{q = 0.052}$) and weakened after excluding the extreme 2R8S target ($\boldsymbol{r = -0.542}$, $\boldsymbol{p = 0.056}$). **Conclusions:** These results indicate that performance on conventional structure-prediction benchmarks does not necessarily translate into improved performance as an RL reward oracle. Reward-landscape disagreement is strongly target-dependent in both reward magnitude and candidate ranking. Robust dispersion measures identify a suggestive association between greater AF3 reward spread and poorer AF3-relative RL performance, but the evidence remains exploratory and stochastic repeatability is not measured directly. More broadly, oracle selection for RL-based RNA design may need to consider operating regime and reward behaviour in addition to structure-prediction accuracy.

Keywords: RNA inverse design, reinforcement learning, reward oracle, RhoFold, AlphaFold3

1 Background

RNA structure is closely connected to molecular function, making computational design of sequences that adopt a desired structure an important problem in synthetic biology and RNA engineering. RNA inverse design, or inverse folding, reverses the conventional structure-prediction problem: rather than predicting structure from a given sequence, it seeks one or more nucleotide sequences compatible with a specified structural target.

1.1 RNA Inverse Design

RNA inverse design encompasses related but distinct problems depending on the structural representation used as the design target. In *secondary-structure inverse folding*, the objective is to identify an RNA sequence predicted to adopt a specified base-pairing pattern, commonly represented in dot-bracket notation. Classical methods such as RNAinverse and INFO-RNA approach this problem through thermodynamic folding models and sequence-space search, while later methods including RNAiFold and antaRNA introduced constraint programming and alternative optimization strategies [1–4]. Reinforcement-learning methods have also been developed for this setting; for example, LEARNA constructs sequences sequentially using feedback from predicted secondary-structure agreement [5]. These approaches demonstrate the use of folding-based feedback for sequence optimization, but their design objective remains a two-dimensional base-pairing topology rather than a complete RNA tertiary structure. Three-dimensional RNA inverse design is a distinct and more recent problem in which the input is an experimentally determined or computationally specified RNA backbone and the objective is to generate sequences compatible with that 3D geometry. Recent deep-learning approaches differ substantially in how they represent the target structure and generate candidate sequences. gRNAde uses a geometric graph neural network (GNN) to encode one or more RNA backbone conformations and generates sequences autoregressively, conditioning each nucleotide prediction on the target geometry and previously generated residues [**?**]. Relational extensions of this framework introduce separate edge types representing backbone connectivity, secondary-structure interactions, and spatial relationships, providing the network with additional structural information beyond geometric proximity alone [6]. Other methods have explored alternative representations and generative strategies for the same 3D design problem. SCRU-Seq and SCRU-Diff operate on structural units derived from RNA 3D data using a dual-radius graph representation intended to capture both local chemical environments and longer-range structural relationships [7]. SCRU-Seq predicts an RNA sequence directly in a non-autoregressive forward pass, whereas SCRU-Diff uses iterative discrete diffusion to generate diverse sequences for a target backbone. RiboDiffusion similarly formulates 3D inverse folding as conditional diffusion, combining a geometric structure encoder with Transformer-based iterative sequence refinement [8]. Together, these methods demonstrate an advance from secondary-structure inverse folding toward generative models conditioned directly on tertiary structural information. Despite sharing a 3D structural input, these approaches also differ in their optimization objectives. Many structure-conditioned models are trained primarily through native sequence recovery, which measures similarity between a generated sequence and the experimentally observed sequence associated with a target backbone. However, a single tertiary structure may be compatible with multiple sequences, so native sequence recovery is only an indirect measure of whether a generated sequence reproduces the intended fold. RIDER addresses this distinction explicitly by combining a 3D structure-conditioned diffusion model with reinforcement-learning fine-tuning [9]. A geometric encoder first represents the target RNA backbone and conditions diffusion-based sequence generation; the resulting policy is then fine-tuned using structural self-consistency between the target structure and the predicted 3D fold of generated sequences. This reinforcement-learning stage introduces a dependency that is absent, or less direct, in purely supervised sequence-recovery approaches. The structure predictor used during RIDER fine-tuning determines the reward assigned to each candidate sequence and therefore defines the reward landscape over which the policy is optimized. Consequently, the properties required of a useful RL folding oracle may differ from those required

of a predictor evaluated only on conventional structure-prediction benchmarks. This motivates the present study of whether replacing RIDER's RNA-specialist structure-prediction oracle with a more general biomolecular structure predictor changes the behaviour and outcome of RL-based RNA inverse design.

1.2 RNA Structure Prediction and Folding Oracles

The distinction between secondary- and tertiary-structure design is also important when considering the prediction models used to evaluate candidate sequences. Secondary-structure prediction methods estimate base-pairing relationships from sequence, typically represented as dot-bracket structures or pair-probability maps. Thermodynamic methods such as RNAfold and Mfold use nearest-neighbour free-energy models, while comparative methods can additionally exploit covariation across homologous sequences. Although these approaches are central to secondary-structure inverse folding, they do not directly provide the three-dimensional coordinates required by RIDER's tertiary-structure reward. Three-dimensional structure predictors instead map RNA sequences to spatial molecular structures that can be compared directly with the target backbone. Recent learned predictors include RNA-specific models such as RhoFold/RhoFold+ and general biomolecular predictors such as AlphaFold3. These models differ not only in architecture and training data, but also in their operating regimes and structure-generation procedures. Such differences are particularly relevant when the predictor is embedded inside an RL loop, where repeated oracle evaluations determine the reward assigned to generated sequences. For an RL-based design framework, prediction accuracy alone therefore does not fully characterize oracle quality. The oracle must produce a reward landscape that provides useful discrimination among candidate sequences and a sufficiently consistent optimization signal for policy updates. Two predictors may achieve similar performance when evaluated against experimentally determined structures while ranking candidate sequences differently within the region of sequence space explored by the design policy. Oracle selection should therefore be considered as an optimization problem as well as a structure-prediction problem. Deep learning methods have recently achieved substantial advances in RNA 3D structure prediction. RhoFold [10] introduced an architecture for RNA tertiary-structure prediction using multiple sequence alignment (MSA)-based inference, and the later RhoFold+ model [11] incorporates RNA-FM, a language model pretrained on RNA sequences, alongside MSA-derived information. AlphaFold3 (AF3) [12] extends the AlphaFold framework to multiple biomolecular modalities, including RNA, and generates structures through a diffusion-based module. Independent benchmarking does not establish a uniform accuracy ordering between AF3 and RNA-specialist predictors: one third-party evaluation reported higher mean TM-score for AF3 than RhoFold+ on an RNA monomer benchmark [13], whereas the RhoFold+ evaluation reports stronger performance for RhoFold+ on its single-RNA comparisons [11]. The two predictors also differ in their standard operating pipelines, although both support documented inference without coevolutionary MSA information. RhoFold exposes a single-sequence inference option (`--single_seq_pred`), while the official AF3 input specification defines RNA MSA-free inference by supplying an empty `unpairedMsa` field, which prevents MSA construction for that RNA chain. In the RL setting studied here, repeated database searches for every generated trajectory are computationally impractical, so both oracle conditions are evaluated in their documented MSA-free modes. The auxiliary analysis in Appendix A therefore serves only to contextualize how AF3 prediction quality changes relative to an MSA-enabled workflow; it is not required to establish the fairness of the head-to-head MSA-free oracle comparison. A second distinction is structure generation: the RhoFold oracle used in the RIDER implementation is deterministic under a fixed configuration, whereas AF3 employs diffusion-based structure generation. These differences motivate examining oracle behaviour inside the RL loop rather than inferring oracle utility solely from conventional prediction benchmarks.

1.3 RIDER: RL-Based RNA Inverse Design

RIDER [9] utilizes a GVP-GNN structure encoder and a variance-preserving diffusion decoder, operating on a 3-bead coarse-grained backbone representation (P, C4′, N1/N9) with a k-nearest-neighbour graph ($k = 32$). Sequences are generated via DDIM-sampled denoising diffusion (30 steps during RL, 50 at inference). Training proceeds in two stages: a supervised RIDE pretraining stage that recovers native sequences from crystallographic structure, followed by RL fine-tuning against a specific target structure via policy gradient updates. For each generated sequence, the folding oracle predicts a 3D structure and computes a self-consistency reward:

$$R = -\left(\tfrac{1}{2}\,\mathrm{RMSD}\right)^2 + \left(5\,\mathrm{GDT\_TS}\right)^2 + \text{bonus}. \tag{1}$$

Here, R is the scalar reward assigned to a generated candidate sequence. RMSD is the root-mean-square deviation, in Å, between corresponding coordinates in the oracle-predicted and target structures after optimal superposition; lower RMSD indicates closer structural agreement, so its squared contribution enters the reward negatively. The factor 1/2 is a fixed scaling coefficient inherited from the RIDER reward implementation. GDT_TS is the Global Distance Test Total Score, a unitless measure of the fraction of the structure that can be superimposed within predefined distance thresholds; higher values indicate better structural agreement. The factor 5 is the corresponding fixed scaling coefficient before squaring. The *bonus* term consists of two mutually exclusive, threshold-triggered linear components, evaluated in priority order: if GDT_TS > 0.45, a bonus of $(\mathrm{GDT\_TS} - 0.45) \times 100$ is added; otherwise, if RMSD < 3Å, a bonus of $(3 - \mathrm{RMSD}) \times 20$ is added. At most one bonus component applies per evaluation. This adds an additional linear contribution after a high-similarity threshold is reached, increasing reward separation among candidates in that regime. All reward coefficients and thresholds are held fixed across oracle conditions, so the folding oracle is the experimental variable.

Crucially, the RL stage fine-tunes on a single fixed target structure at a time, making each run a probe of the oracle's reward signal for that specific target. The oracle therefore has two related roles: it defines the reward landscape the policy navigates, and it supplies the training signal that shapes subsequent sequence generation. These properties are distinct from, and not guaranteed by, raw prediction accuracy measured outside the RL setting. Because the oracle maps each candidate sequence to a predicted structure and resulting reward, it can in principle be substituted for another structure-prediction model exposing the same interface. We use this property to compare the native RNA-specialist oracle RhoFold with AF3 while holding the remaining RIDER training pipeline fixed.

1.4 Study Objective and Hypotheses

We investigate whether replacing RIDER's RhoFold folding oracle with AF3 changes RL optimization behaviour and the structural self-consistency reached under otherwise fixed training conditions. Beyond measuring the direct performance difference between oracle conditions, we consider two hypotheses that could account for an observed gap. **H1 (reward-landscape alignment)** proposes that oracle utility depends partly on how the reward landscape induced by a predictor aligns with the PDB-derived target structures and structural metrics used by RIDER; two predictors can therefore rank the same candidate sequences differently even when both are accurate structure predictors. **H2 (reward dispersion and consistency)** distinguishes two related properties of the oracle signal: dispersion of rewards across different candidate sequences and stochastic repeatability when the same sequence is evaluated repeatedly. Either property could affect PPO optimization by changing the distribution or uncertainty of advantage estimates used for policy updates. The present study directly measures across-sequence reward dispersion and training dynamics; repeated folds of identical sequences would be required to isolate stochastic oracle repeatability. These hypotheses motivate the reward-distribution, training-dynamics, and auxiliary MSA-sensitivity analyses presented below; they are evaluated as possible explanatory mechanisms rather than assumed to be the causes of any observed oracle-performance difference.

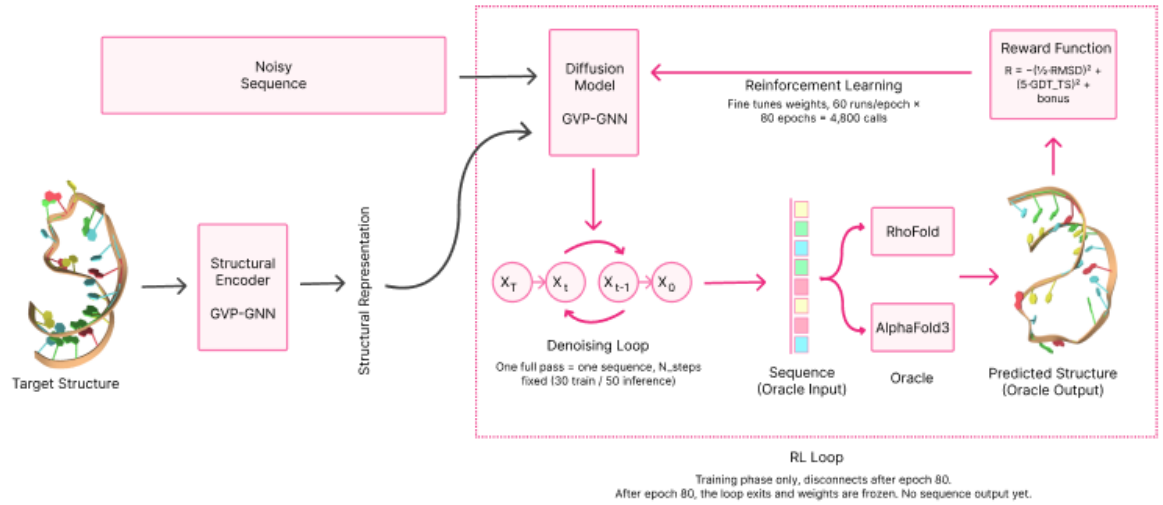

Fig. 1 RIDER RL fine-tuning with a swappable folding oracle. Adapted from Hu et al. [9]. The diffusion model generates a candidate sequence via DDIM denoising, which is folded by the swappable oracle (RhoFold or AlphaFold3) and scored against the target structure using the reward function (Equation 1). The oracle substitution investigated in this paper occurs at this step.

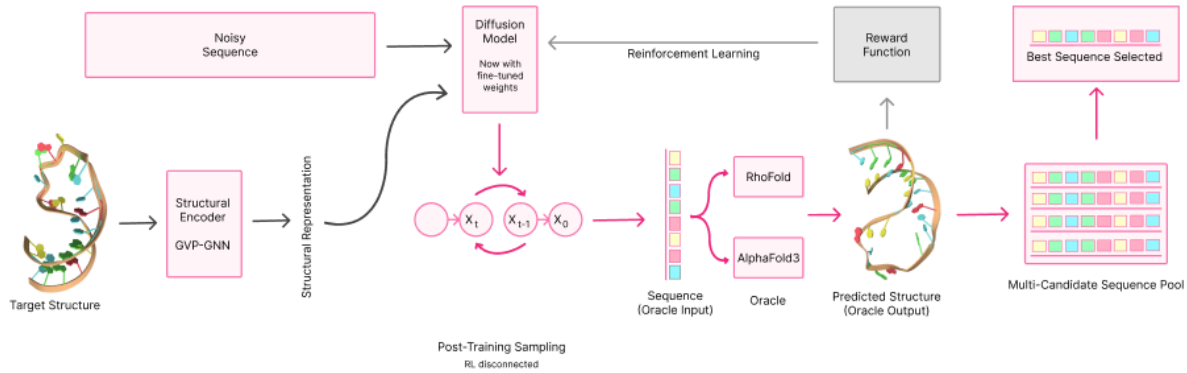

Fig. 2 Post-training sampling after RIDER fine-tuning. Adapted from Hu et al. [9]. Once RL fine-tuning ends, the reinforcement-learning feedback loop and reward function are disconnected (greyed); the diffusion model, now holding fine-tuned weights, generates a pool of candidate sequences via the same denoising process used during training. Each candidate is independently folded by the oracle and scored for self-consistency against the target structure; the highest-scoring candidate is selected as the final design.

2 Methods

2.1 Oracle Configurations

RhoFold is used as the RNA-specialist oracle in the comparison. The experimental condition uses the checkpoint `model_20221010_params.pt` [10], with documented single-sequence inference enabled (`--single_seq_pred True`). This checkpoint-and-mode combination defines the RhoFold condition reported throughout the manuscript; RhoFold+ is discussed separately only when citing later structure-prediction literature. **AlphaFold3** [12] is run via Apptainer in the documented RNA MSA-free operating mode: the RNA `unpairedMsa` input is set to an empty string and the data-pipeline search is disabled (`--run_data_pipeline=false`). The official AF3 input specification defines an empty RNA `unpairedMsa` as MSA-free inference. This mode is

used because database search for each of the 60 sequences generated per RL epoch is computationally impractical inside the RL loop. Both oracles are accessed through a common wrapper interface (oracle.py), with a single fold() method regardless of the underlying model, so the RL loop's sampling and reward-computation code is identical across oracle conditions. Which oracle is active for a given run is selected via a single configuration field (config.oracle, set to rhofold or alphafold3 in the run's YAML configuration) and instantiated through function (get_oracle()). This design is what makes the oracle substitution investigated in this paper a controlled swap.

2.2 RL Hyperparameters

All 28 runs (14 structures $\times$ 2 oracles) share the hyperparameters in Table 1 and the same pre-trained RIDE checkpoint.

2.3 Experimental Design

We hold the RIDER architecture, pre-trained RIDE checkpoint, and all RL hyperparameters fixed and vary only the folding oracle, isolating the oracle's contribution to performance. We align our oracle comparison with the standard benchmark established by Das et al. [14] and adopted by gRNAde [**?**] and RIDER [9], comprising 14 RNA structures spanning riboswitches, aptamers, ribozymes, and structural RNA elements (Table 2), all held out from training.

2.4 Reward Distribution Sampling Protocol

To characterize each oracle's reward signal independent of RL training dynamics, the reward-distribution comparison (Section 3.3) was extended from a single preliminary target to all 14 Das14 benchmark structures. For each target, a fixed pool of 100 sequences was sampled from the pre-trained RIDE checkpoint before any RL fine-tuning. Each candidate sequence was then folded by both oracles and scored with the RIDER reward function (Equation 1) against the corresponding target structure, yielding the per-target statistics reported in Table 5. Sequences are generated using DDIM sampling at temperature 0.5 with 30 denoising steps from the pre-trained RIDE checkpoint, representing the policy's prior distribution with no influence from either oracle. To distinguish variation across candidate sequences from stochastic AF3 prediction variability, repeated AF3 predictions of identical sequences across multiple random seeds should be analysed separately in future work (Section 4.3).

2.5 Reward-Distribution Statistical Analysis

For each target, the primary agreement statistic is the Pearson correlation coefficient r, computed on the 100 paired (RhoFold, AF3) rewards assigned to the same candidate sequences. Pearson r measures the strength and direction of a linear relationship between the two reward scales, with $r = 1$ denoting perfect positive linear agreement, $r = 0$ no linear association, and $r = -1$ perfect inverse linear association. The associated two-sided p value tests the null hypothesis $r = 0$; because 14 target-level tests are performed, $p < 0.05$ is described as nominal significance rather than family-wise evidence. Benjamini–Hochberg false-discovery-rate (FDR) correction is applied separately to the 14 Pearson and 14 Spearman target-level tests at $q = 0.05$.

Spearman's rank correlation coefficient ρ is computed on the same paired rewards as a rank-based sensitivity analysis. Spearman ρ measures monotonic agreement in candidate ordering rather than requiring a linear relationship on the original reward scale. It is particularly relevant to H1 because PPO optimization depends on relative reward ordering as well as reward magnitude.

Reward spread is characterized by three AF3/RhoFold dispersion ratios. The conventional variance ratio is $\sigma^2_{\mathrm{AF3}}/\sigma^2_{\mathrm{RhoFold}}$. Two robust sensitivity measures are also computed: the median-absolute-deviation ratio, $\mathrm{MAD}_{\mathrm{AF3}}/\mathrm{MAD}_{\mathrm{RhoFold}}$, where MAD is the median absolute distance from the sample median; and the interquartile-range ratio, $\mathrm{IQR}_{\mathrm{AF3}}/\mathrm{IQR}_{\mathrm{RhoFold}}$, where IQR is the difference between the 75th and 25th percentiles. Ratios above 1 indicate broader AF3 reward dispersion and ratios below 1 indicate broader RhoFold dispersion. MAD and IQR reduce the influence of extreme reward observations relative to variance.

For cross-target associations with RL outcome (Section 3.3), each dispersion ratio is natural-log

Table 1 RL hyperparameters held fixed across all oracle conditions.

Hyperparameter	Value
RL epochs	80
Learning rate	5×10^{-5}
Trajectories per epoch	60
Policy-gradient updates per epoch	2
PPO clip ε	0.5
Baseline EMA β	0.8
DDIM steps (train / inference)	30 / 50
Sampling temp. (train / eval)	0.5 / 0.1
Precision	fp16
Oracle-specific settings	
RhoFold checkpoint/mode	`model_20221010_params.pt`; single-seq. (`--single_seq_pred True`)
AF3 mode	RNA MSA-free (empty `unpairedMsa`; `run_data_pipeline=false`)

Table 2 The 14 RNA structures of interest identified by Das et al. [14], used as the benchmark for oracle comparison in this work. Sample indices refer to positions in the DAS test split of the gRNAde / RIDER dataset. Lengths reflect the number of nucleotides after masking non-standard residues.

PDB	Description	Length (nt)	Class	Sample idx
1X9C	All-RNA hairpin ribozyme	60	Ribozyme	1
4FE5	Guanine riboswitch aptamer	67	Riboswitch	19
2GDI	Thiamine pyrophosphate riboswitch	78	Riboswitch	52
1L2X	Viral RNA pseudoknot	27	Pseudoknot	62
2GCS	*glmS* ribozyme (pre-cleavage)	122	Ribozyme	66
1ET4	Vitamin B12 binding RNA aptamer	35	Aptamer	75
1Q9A	Sarcin/ricin domain, 23S rRNA	27	Structural RNA	76
2R8S	FAB-bound ribozyme domain	159	Ribozyme	84
1XPE	HIV-1 RNA dimerization site	46	Structural RNA	91
1CSL	RRE high affinity site	28	Aptamer	93
1LNT	RNA internal loop of SRP	22	Structural RNA	94
1F27	Biotin-binding RNA pseudoknot	19	Pseudoknot	95
2OEU	Junctionless hairpin ribozyme	42	Ribozyme	96
354D	Loop E from *E. coli* 5S rRNA	23	Structural RNA	97

transformed before correlation. The log transformation maps equal dispersion to 0, AF3-greater dispersion to positive values, and RhoFold-greater dispersion to negative values while reducing right skew. Δ_{RL} is the outcome variable reported in Table 3, defined as AF3 peak TM minus RhoFold peak TM; negative values therefore indicate a RhoFold advantage. We correlate per-target Pearson r and Spearman ρ with Δ_{RL} to test whether oracle agreement predicts downstream oracle-relative performance. We separately correlate the log variance, log MAD, and log IQR ratios with Δ_{RL}. Because these three dispersion–outcome associations are exploratory tests of the same mechanism, their p values are additionally adjusted together using Benjamini–Hochberg FDR correction. Sensitivity to the extreme 2R8S dispersion target is assessed by repeating the dispersion–outcome correlations after excluding that target.

3 Results

3.1 AF3 Shows Lower RL Oracle Performance Than RhoFold on Most Targets

We test RL oracle performance by comparing RhoFold and AF3 under otherwise fixed RIDER settings. Table 3 reports oracle-specific peak TM-score and minimum RMSD across all 14 structures. RhoFold achieved a higher peak TM-score on **10 of 14 structures**, with a mean peak TM-score advantage of 0.105 (0.437 vs. 0.332). AF3 achieved a higher peak TM-score on three targets (1LNT, 354D, and 1Q9A), while 1XPE was effectively tied ($\Delta_{\mathrm{RL}} = +0.001$). The largest RhoFold advantages occur among several longer ribozyme and riboswitch targets. Across the six ribozyme or riboswitch targets of length 42–159 nt, RhoFold achieved a mean peak TM-score of 0.618 compared with 0.408 for AF3. By contrast, AF3 achieved higher peak TM-scores on three short structural RNAs ($\leq$27 nt), with 1XPE effectively tied.

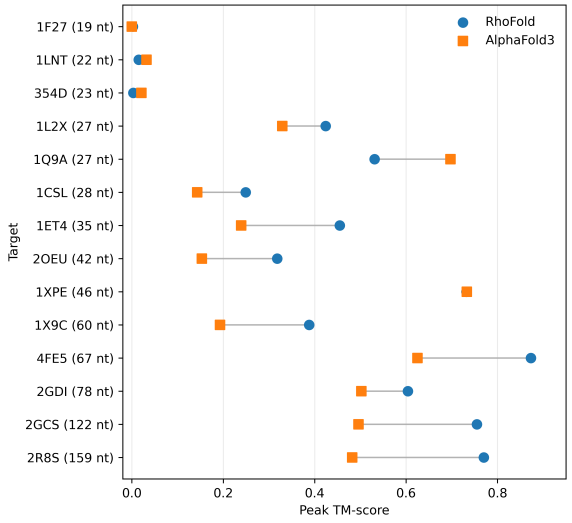

Fig. 3 Peak TM-score across Das14 targets. Connected markers compare RhoFold and AF3 oracle conditions for each target.

3.2 Training Dynamics

Plateau TM is the mean TM-score over the final 20 logged steps; *Collapse %* is the fraction of post-warmup logged RL steps (logged step > 50) with TM < 0.1 (Table 4). Collapse rates were not uniform across the benchmark. Both oracles showed 100% collapse on the three shortest structures (1F27, 1LNT, 354D), while both showed 0% collapse on 1L2X, 1Q9A, 1XPE, and 4FE5. Among the remaining seven targets, AF3 showed a higher collapse rate than RhoFold on six (1CSL, 1ET4, 2OEU, 1X9C, 2GCS, and 2R8S); 2GDI was the exception, with collapse rates of 0.9% for AF3 and 1.9% for RhoFold. The largest difference occurred on 2OEU (51.7% vs. 21.2%).

3.3 Oracle Reward Distribution Comparison

To characterize each oracle's reward signal independent of RL training dynamics, we extended the reward-distribution comparison to all 14 Das14 benchmark targets. For each target, a fixed pool of 100 sequences was sampled from the pretrained RIDE checkpoint before any RL fine-tuning, folded by both oracles, and scored with the RIDER reward function (Equation 1) against the corresponding target structure (Section 2.4).

Across targets, reward agreement was generally weak in both magnitude and ranking. Mean Pearson correlation was $r = 0.103$ (median 0.091), while mean Spearman rank correlation was $\rho = 0.069$ (median 0.085). Five targets showed nominal Pearson $p < 0.05$ (2GDI, 1L2X, 1ET4, 1Q9A, 1LNT), and four showed nominal Spearman $p < 0.05$ (1LNT, 1L2X, 1Q9A, 1ET4). After Benjamini–Hochberg correction, the same three targets remained significant for both statistics: 1LNT ($r = 0.528$, $\rho = 0.630$), 1L2X ($r = 0.426$, $\rho = 0.425$), and 1Q9A ($r = -0.332$, $\rho = -0.738$). The strong negative rank correlation on 1Q9A indicates that oracle disagreement extends beyond reward scaling: the two oracles preferentially rank candidate sequences in substantially opposing orders on this target.

Reward dispersion was also highly target-dependent. The mean AF3/RhoFold variance ratio was 2.828, but the median was 0.980, with AF3 showing lower variance on 7 of 14 targets. Robust measures showed the same heterogeneity: mean MAD and IQR ratios were 1.895 and 2.014, whereas the corresponding medians were 0.800 and 0.856; AF3 showed lower MAD and IQR dispersion on 8 of 14 targets. Thus, mean ratios above 1 do not indicate uniformly broader AF3 rewards.

Table 3 Peak oracle-specific structural self-consistency (TM-score) and minimum RMSD achieved during RL fine-tuning, comparing RhoFold and AF3 (no-MSA) as folding oracles. Δ_{RL} = AF3 peak TM − RhoFold peak TM; negative values indicate a higher peak TM-score for RhoFold. Bold indicates the better-performing oracle per structure.

PDB	Class	nt	RF TM	AF3 TM	Δ_{RL} TM	RF RMSD	AF3 RMSD
1F27	Pseudoknot	19	**0.002**	0.000	−0.002	**1.745**	3.450
1LNT	Structural RNA	22	0.015	**0.032**	+0.017	**3.030**	3.473
354D	Structural RNA	23	0.003	**0.021**	+0.018	**4.388**	12.965
1L2X	Pseudoknot	27	**0.424**	0.329	−0.095	**1.618**	2.081
1Q9A	Structural RNA	27	0.531	**0.697**	+0.166	**1.003**	0.682
1CSL	Aptamer	28	**0.249**	0.143	−0.106	**2.584**	3.124
1ET4	Aptamer	35	**0.455**	0.239	−0.217	**2.218**	4.195
2OEU	Ribozyme	42	**0.318**	0.153	−0.165	**6.366**	8.243
1XPE	Structural RNA	46	**0.732**	**0.733**	+0.001	**1.707**	1.731
1X9C	Ribozyme	60	**0.388**	0.193	−0.195	**4.687**	7.630
4FE5	Riboswitch	67	**0.873**	0.625	−0.248	**1.152**	2.438
2GDI	Riboswitch	78	**0.604**	0.502	−0.102	**3.118**	4.599
2GCS	Ribozyme	122	**0.755**	0.496	−0.258	**2.817**	8.951
2R8S	Ribozyme	159	**0.770**	0.482	−0.289	**3.334**	8.825
Mean			**0.437**	0.332	−0.105	**3.112**	5.099

Table 4 Training-dynamics summary for all 28 Das14 RL runs (14 structures × 2 oracles).

PDB	Class	nt	RF Plateau TM	AF3 Plateau TM	RF Collapse %	AF3 Collapse %
1F27	Pseudoknot	19	0.000	0.000	100.0%	100.0%
1LNT	Structural RNA	22	0.014	**0.032**	100.0%	100.0%
354D	Structural RNA	23	0.003	**0.021**	100.0%	100.0%
1L2X	Pseudoknot	27	**0.416**	0.322	0.0%	0.0%
1Q9A	Structural RNA	27	0.531	**0.697**	0.0%	0.0%
1CSL	Aptamer	28	**0.249**	0.136	15.9%	26.8%
1ET4	Aptamer	35	**0.453**	0.202	0.5%	2.6%
2OEU	Ribozyme	42	**0.318**	0.146	21.2%	51.7%
1XPE	Structural RNA	46	0.713	**0.720**	0.0%	0.0%
1X9C	Ribozyme	60	**0.333**	0.172	7.0%	31.2%
4FE5	Riboswitch	67	**0.873**	0.582	0.0%	0.0%
2GDI	Riboswitch	78	**0.562**	0.502	1.9%	0.9%
2GCS	Ribozyme	122	**0.753**	0.492	0.0%	0.2%
2R8S	Ribozyme	159	**0.754**	0.412	0.0%	0.9%
Mean			**0.427**	0.317	24.7%	29.6%

The largest exception was 2R8S, which remained extreme under variance (25.43), MAD (11.18), and IQR (11.30), showing that its increased dispersion is not solely an artifact of variance sensitivity to isolated outliers. 354D and 1ET4 also showed consistently elevated AF3 dispersion across the robust measures.

Neither target-level Pearson reward alignment nor rank alignment explained the downstream oracle-performance difference. Correlating per-target Pearson r with Δ_{RL} gave $r = -0.050$ ($p = 0.865$), while correlating Spearman ρ with Δ_{RL} gave $r = -0.278$ ($p = 0.336$). Among the three dispersion measures, the log MAD ratio showed the strongest exploratory association with Δ_{RL}

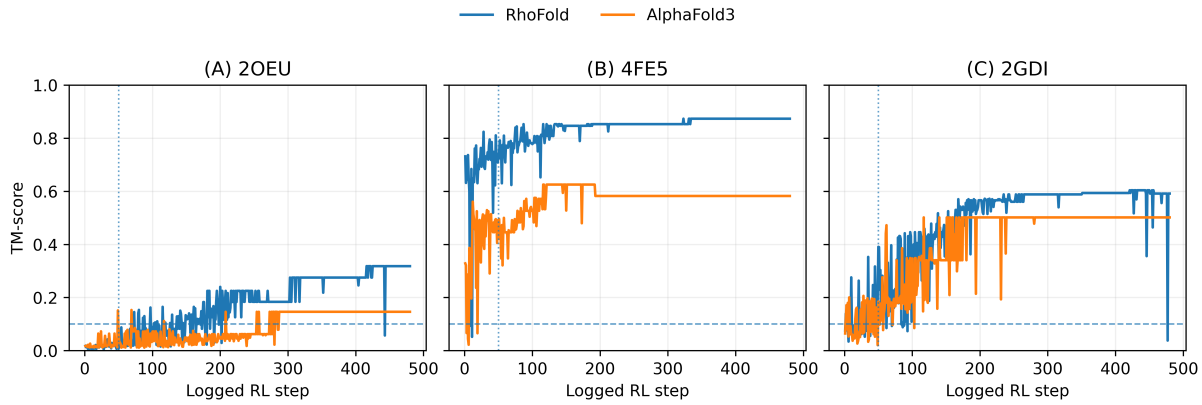

Fig. 4 Representative RL training dynamics under RhoFold and AF3. TM-score trajectories are shown for 2OEU, 4FE5, and 2GDI to illustrate distinct target-level training behaviours.

Table 5 RhoFold versus AlphaFold3 reward-alignment and dispersion statistics across the full 14-structure Das et al. benchmark. For each target, 100 identical candidate sequences sampled from the pre-trained checkpoint were scored by both oracles. Pearson r measures linear reward agreement and Spearman ρ measures rank-order agreement. Dispersion ratios are AF3/RhoFold; values above 1 indicate broader AF3 rewards. Bold correlation coefficients survive Benjamini–Hochberg correction at $q < 0.05$ within the corresponding 14 target-level tests.

PDB	Pearson r	Spearman ρ	Var. ratio	MAD ratio	IQR ratio
1X9C	0.128	0.079	1.423	0.859	0.847
4FE5	−0.002	0.099	0.537	1.543	1.381
2GDI	0.227	0.177	2.482	1.029	1.250
1L2X	**0.426**	**0.425**	1.633	1.240	1.487
2GCS	0.009	−0.068	1.748	0.652	0.865
1ET4	0.233	0.206	1.805	2.855	2.855
1Q9A	−0.332	−0.738	0.403	0.006	0.275
2R8S	−0.106	−0.018	25.427	11.180	11.301
1XPE	−0.056	−0.093	0.347	0.742	0.604
1CSL	0.061	0.090	0.074	0.092	0.050
1LNT	**0.528**	**0.630**	0.178	0.104	0.216
1F27	0.151	0.026	0.134	0.121	0.101
2OEU	0.056	0.037	0.203	0.123	0.146
354D	0.120	0.115	3.195	5.988	6.820
Mean	0.103	0.069	2.828	1.895	2.014
Median	0.091	0.085	0.980	0.800	0.856

($r = -0.622$, $p = 0.017$), indicating that greater robust AF3 reward dispersion tended to coincide with poorer AF3-relative RL performance. However, this association weakened after excluding 2R8S ($r = -0.542$, $p = 0.056$, $n = 13$) and did not survive Benjamini–Hochberg correction across the variance-, MAD-, and IQR-outcome tests ($q = 0.052$ for MAD). The corresponding full-sample associations for log variance ratio ($r = -0.455$, $p = 0.102$) and log IQR ratio ($r = -0.388$, $p = 0.171$) were weaker. These results make robust reward dispersion a hypothesis-generating signal rather than a confirmed predictor of RL oracle performance.

4 Discussion

Across the Das14 benchmark, AF3 achieved lower peak TM-scores than RhoFold on most targets under the oracle conditions tested here. The

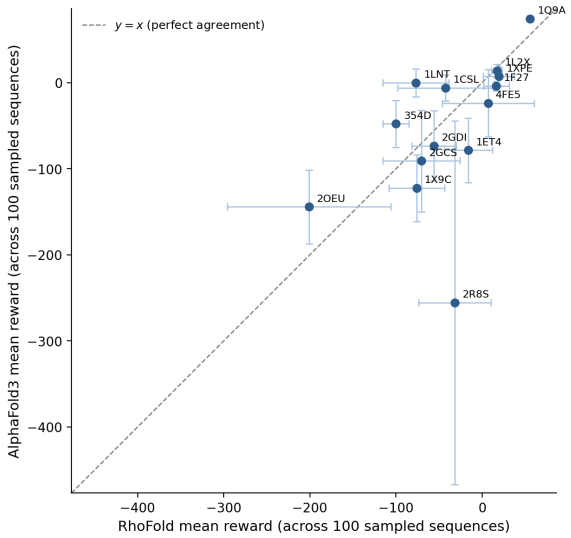

Fig. 5 Oracle mean-reward comparison across Das14 targets. Each point is one target's mean reward ± one standard deviation over 100 candidate sequences sampled from the pre-trained checkpoint; the dashed line marks equal mean reward ($y = x$). Points below the line indicate a lower AF3 mean reward for the same candidate pool. 2R8S is a substantial outlier in both position and dispersion.

comparison therefore indicates that the relative performance of a structure predictor on conventional prediction benchmarks does not directly determine its usefulness as an RL reward oracle.

4.1 H1: Reward-Landscape Alignment

The full-benchmark results are consistent with H1 in showing that the two predictors can assign substantially different rewards to the same candidate sequence space, but the effect is heterogeneous rather than uniform. Agreement was weak on average in both reward magnitude (mean Pearson $r = 0.103$) and candidate ordering (mean Spearman $\rho = 0.069$), and only three of fourteen targets survived multiple-comparison correction for either statistic (1L2X, 1Q9A, 1LNT; Table 5). The rank-based result is particularly informative for 1Q9A: Spearman $\rho = -0.738$ indicates strong inverse candidate ordering, so the disagreement cannot be explained simply by one oracle using a shifted or rescaled reward range.

These results establish that RhoFold and AF3 define substantially different, target-dependent reward landscapes, but they do not identify reward-landscape disagreement itself as the mechanism responsible for RhoFold's stronger RL performance. Across targets, neither Pearson alignment ($r = -0.050$, $p = 0.865$ versus Δ_{RL}) nor Spearman rank alignment ($r = -0.278$, $p = 0.336$) predicts the observed oracle-performance gap. The paired-reward experiment therefore establishes disagreement between the landscapes, but it does not determine which oracle's candidate ranking is better aligned with an independent ground truth, since no oracle-independent structural ranking of the 100-sequence candidate pools was available.

4.2 H2: Reward Dispersion and Consistency

H2 distinguishes *reward dispersion* from *stochastic reward consistency*. Reward dispersion describes how widely an oracle scores different candidate sequences within a target, whereas stochastic consistency describes how reproducibly the oracle scores the same sequence when prediction is repeated. These properties are related but not interchangeable: broad across-sequence dispersion can reflect useful discrimination among candidates, while poor repeatability can introduce noise into PPO advantage estimates. AF3 uses stochastic diffusion-based structure generation, so repeated predictions may vary across model seeds and diffusion samples. The present full-benchmark analysis measures only across-sequence dispersion.

The robust sensitivity analyses clarify that broader AF3 dispersion is not a uniform benchmark-wide property. Although the mean AF3/RhoFold variance, MAD, and IQR ratios are 2.83, 1.90, and 2.01, their medians are 0.98, 0.80, and 0.86, respectively. AF3 has lower variance on 7 of 14 targets and lower MAD/IQR dispersion on 8 of 14. At the same time, 2R8S remains extreme under all three measures (25.43, 11.18, and 11.30), and 354D remains elevated under both robust measures, showing that the high-dispersion behaviour on these targets is not merely an artifact of variance sensitivity to isolated outliers.

The strongest exploratory relationship with downstream RL outcome is obtained with robust dispersion: the log MAD ratio correlates with Δ_{RL} at $r = -0.622$ ($p = 0.017$), so greater AF3-relative MAD tends to occur on targets where AF3 performs worse relative to RhoFold. However, the result is not confirmatory. It weakens

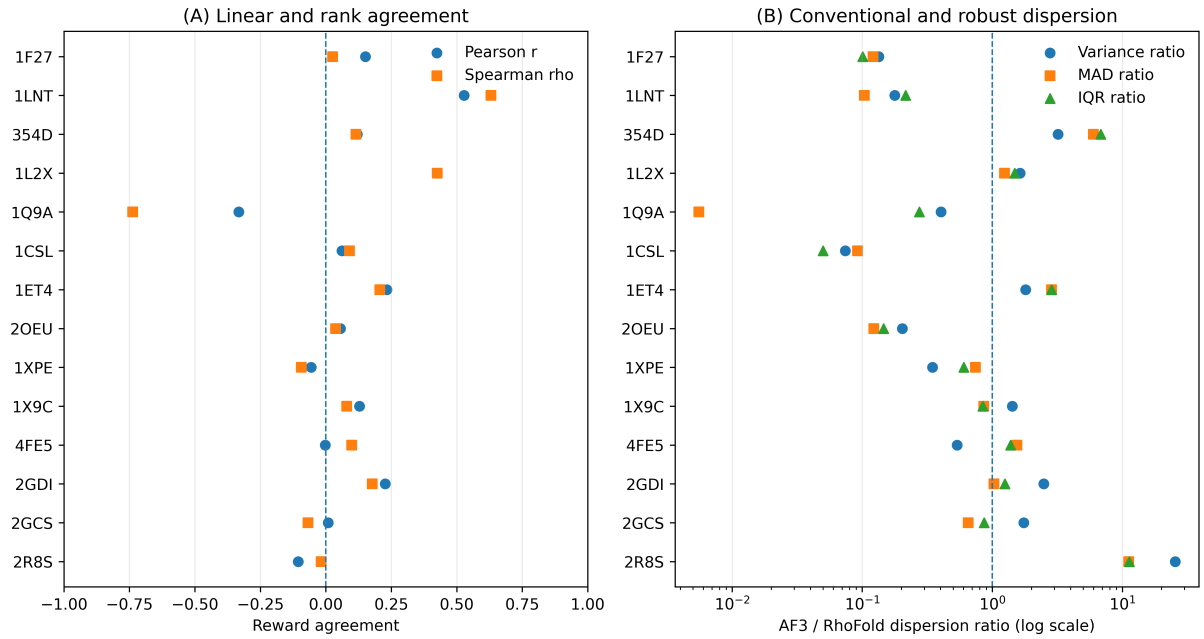

Fig. 6 Reward-agreement and dispersion sensitivity analyses. (A) Target-level Pearson r and Spearman ρ compare linear reward agreement with rank-order agreement for the same paired candidate sequences. (B) Variance, MAD, and IQR ratios compare conventional and robust measures of AF3-to-RhoFold reward dispersion; a ratio of 1 denotes equal dispersion. The persistence of the 2R8S and 354D elevations under MAD and IQR indicates that their broader AF3 reward distributions are not explained solely by variance sensitivity to isolated extreme observations.

after excluding 2R8S ($r = -0.542$, $p = 0.056$) and narrowly misses FDR significance when corrected across the three dispersion–outcome tests ($q = 0.052$). The log variance and log IQR associations are weaker ($r = -0.455$, $p = 0.102$ and $r = -0.388$, $p = 0.171$, respectively). The training-dynamics results provide separate target-level evidence that optimization behaviour differs between oracle conditions: among the seven targets with non-identical, nonzero collapse behaviour, AF3 shows a higher collapse rate on six, with 2GDI as the exception (Section 3.2). Taken together, across-sequence dispersion remains a plausible contributor to oracle behaviour, but the evidence is exploratory, and repeated predictions of identical sequences across multiple AF3 seeds are still required to evaluate stochastic reward repeatability directly.

Because both RhoFold and AF3 support documented MSA-free inference, the auxiliary AF3 MSA-sensitivity analysis is not needed to justify the head-to-head oracle comparison; rather, it provides context for how AF3 behaves when moved from its usual MSA-enabled workflow into the computationally constrained RL operating regime. Across the 13 structures for which this comparison is available, the association between Δ_{MSA} and Δ_{RL} is weak and not statistically significant ($r = 0.207$, $p = 0.50$). Several targets show little change in AF3 TM-score between the two workflows while still showing a RhoFold advantage during RL, including 1ET4 ($\Delta_{\mathrm{MSA}} = -0.014$), 2OEU ($-0.004$), 1X9C ($+0.001$), and 1CSL ($-0.048$). However, the MSA-enabled predictions were obtained from the AlphaFold Server whereas the MSA-free predictions used the local Apptainer deployment, so this analysis is not a controlled MSA-only ablation and should be interpreted as contextual sensitivity evidence rather than a causal estimate of the MSA contribution. These findings motivate several considerations for oracle selection in RL-based biomolecular design beyond raw prediction accuracy. In particular, reward consistency, reward-landscape alignment, and compatibility with the operating regime may affect optimization behaviour, with the present benchmark-wide analysis providing only exploratory evidence regarding reward dispersion and consistency. The present results

suggest that an RNA-specialist oracle can provide a more effective reward signal than a general biomolecular predictor in this RIDER setting, even when conventional structure-prediction benchmarks do not show a uniform accuracy advantage for the specialist model.

4.3 Limitations

Self-consistency evaluation. Structural fidelity in Table 3 is measured via self-consistency: the active oracle for a given run both provides the RL training signal and evaluates the resulting candidate sequences, since each generated sequence's TM-score and RMSD are computed from that same oracle's own structural prediction. This design means the reported oracle comparison partially conflates two effects: differences in the RL policy's sequence-generation quality, and differences in each oracle's own folding accuracy under the shared no-MSA operating constraint. Cross-oracle evaluation, scoring RhoFold-trained candidates with AF3 and vice versa, would be needed to separate these two effects and is not performed in the present study. All AF3 oracle runs use the documented RNA MSA-free operating mode, implemented locally by supplying an empty RNA MSA input and disabling data-pipeline search; the comparison therefore characterizes AF3 specifically in this computationally constrained operating regime. The current oracle–target comparison uses a single RL run per condition, so run-to-run variation in RL optimization is not yet quantified; additional seeded runs would strengthen estimates of the robustness of the observed oracle differences. Separately, the 14-structure Das et al. benchmark, while standard for RIDER and gRNAde evaluation, is substantially smaller than the 100+ structure benchmarks increasingly used elsewhere in the RNA structure-prediction literature; the time frame of the present oracle comparison (28 separate single-target training runs) made benchmark expansion computationally infeasible within the scope of this study, and the reported findings should be read as characterizing this specific, well-established but modestly sized benchmark rather than establishing generalization across the full diversity of RNA structural space. Separately, repeated AF3 folds of identical sequences across multiple prediction seeds are needed to quantify stochastic oracle variability directly, distinct from the across-sequence variance reported in Section 3.3. The dispersion–outcome analyses reported in Section 3.3 are based on only $n = 14$ targets. The strongest relationship, for log MAD ratio, has nominal $p = 0.017$ but does not survive correction across the three exploratory dispersion measures ($q = 0.052$) and weakens to $p = 0.056$ after excluding 2R8S; it should therefore be treated as hypothesis-generating rather than confirmatory. On the three hardest targets (1F27, 1LNT, 354D), neither oracle achieves meaningful structural fidelity. The comparison is limited to RhoFold and AF3; other single-sequence predictors were considered but deemed impractical as RL oracles.

5 Conclusions

Under the conditions tested in RIDER, substituting AF3 for RhoFold as the RL reward oracle was associated with lower peak oracle-specific structural self-consistency on most Das et al. benchmark targets. Because independent structure-prediction benchmarks provide mixed evidence regarding the relative accuracy of AF3 and RhoFold, these results do not support a simple correspondence between conventional predictor accuracy and RL-oracle utility. The auxiliary AF3 MSA-sensitivity analysis is weakly associated with the observed RL differences and therefore does not identify MSA availability as the primary explanation for the benchmark-wide pattern. Reward-landscape agreement between oracles is weak and highly target-dependent in both reward magnitude and candidate ranking. Robust dispersion analysis identifies a stronger, negative association between AF3-relative MAD and AF3-relative RL outcome, but the relationship is sensitive to 2R8S and narrowly misses FDR significance across the three dispersion tests, so it remains exploratory. Stochastic repeatability of identical AF3 sequence evaluations is not measured here. Overall, the results indicate that oracle selection for RL-based RNA design should consider reward behaviour and operating-regime compatibility alongside structure-prediction accuracy.

Appendix A AF3 MSA Sensitivity Analysis

Both RhoFold and AF3 support documented MSA-free RNA inference, so MSA removal is not a methodological asymmetry in the main oracle comparison. This auxiliary analysis instead asks how AF3 prediction quality changes when the model is moved from an MSA-enabled workflow into the MSA-free operating regime required by repeated RL oracle calls. We folded the native sequence of each benchmark structure under an MSA-enabled AF3 workflow and under the local MSA-free configuration used by the RL oracle, then compared predicted C4′ coordinates with the deposited crystal structure outside the RL loop. MSA-enabled predictions were obtained via the AlphaFold Server (alphafoldserver.com); MSA-free predictions used the local Apptainer deployment with an empty RNA `unpairedMsa` input and data-pipeline search disabled. Because deployment pathway and associated server/local settings are not identical between these two conditions, this analysis is not treated as a controlled MSA-only ablation or as a causal estimate of the effect of MSA removal. It is retained as a contextual sensitivity analysis of AF3's operating regime.

Table A1 compares the MSA-enabled and MSA-free scores for 13 analyzable targets. The largest decreases under the MSA-free workflow occur for 2GDI ($\Delta = -0.596$), 2R8S (-0.274), and 2GCS (-0.218), whereas 1L2X ($+0.081$), 1XPE ($+0.050$), and 1X9C ($+0.001$) show no decrease.

Table A1 AF3 sensitivity to MSA-enabled versus MSA-free workflows. Structure-prediction accuracy is shown for 13 of 14 Das et al. targets. Δ_{MSA} = TM(MSA-free) − TM(MSA-enabled); negative values indicate a lower TM-score in the MSA-free workflow. Because the two workflows used server and local deployments, respectively, this table is interpreted as a contextual sensitivity comparison rather than a controlled MSA-only ablation.

PDB	Description	nt	TM MSA-free	TM MSA-enabled	Δ_{MSA} TM	RMSD MSA-free	RMSD MSA-enabled
1F27	Biotin-binding pseudoknot	19	0.003	0.002	+0.001	4.29	4.82
1LNT	Internal loop of SRP	22	0.005	0.004	+0.001	14.75	13.84
1L2X	Viral RNA pseudoknot	27	0.486	0.405	+0.081	1.57	1.69
1Q9A	Sarcin/ricin domain 23S rRNA	27	0.795	0.891	−0.096	0.57	0.40
1CSL	RRE high affinity site	28	0.183	0.231	−0.048	4.25	3.00
1ET4	Vitamin B12 RNA aptamer	35	0.204	0.218	−0.014	8.99	4.79
2OEU	Junctionless hairpin ribozyme	42	0.016	0.020	−0.004	25.44	25.59
1XPE	HIV-1 dimerization site	46	0.454	0.404	+0.050	3.87	3.56
1X9C	All-RNA hairpin ribozyme	60	0.023	0.022	+0.001	21.03	21.43
4FE5	Guanine riboswitch aptamer	67	0.857	0.909	−0.052	1.24	0.98
2GDI	TPP riboswitch	78	0.193	0.789	−0.596	11.23	1.92
2GCS	*glmS* ribozyme	122	0.470	0.688	−0.218	9.18	7.35
2R8S	FAB-bound ribozyme domain	159	0.375	0.648	−0.274	11.88	7.29
Mean			0.313	0.402	−0.089	9.54	7.43

354D (Loop E from *E. coli* 5S rRNA, 23 nt) is excluded from this comparison. The structure contains non-standard nucleotides (modified guanosine G46, deoxyguanosine DG) that cause residue count inconsistencies between the dataset representation (23 nt) and AF3 predictions (21–22 nt), precluding a fair coordinate-level comparison. 354D is retained in all RL fine-tuning experiments where RIDER's internal masking handles non-standard residues consistently.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Code used for the oracle-comparison experiments is available at https://github.com/amandanew220/RIDER/tree/oracle-comparison.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by an NSERC Undergraduate Student Research Award (USRA) awarded to A.N.

Authors' contributions

A.N. designed and conducted the experiments, performed the analyses, and drafted the manuscript. H.H.T. supervised the project and reviewed and edited the manuscript. Both authors read and approved the manuscript.

Acknowledgements

Not applicable.

References

[1] Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie / Chemical Monthly. 1994;125(2):167–188. https://doi.org/10.1007/BF00818163.

[2] Busch A, Backofen R. INFO-RNA—a fast approach to inverse RNA folding. Bioinformatics. 2006;22(15):1823–1831. https://doi.org/10.1093/bioinformatics/btl194.

[3] Garcia-Martin JA, Clote P, Dotu I. RNAiFOLD: a constraint programming algorithm for RNA inverse folding and molecular design. Journal of Bioinformatics and Computational Biology. 2013;11(2):1350001. https://doi.org/10.1142/S0219720013500017.

[4] Kleinkauf R, Mann M, Backofen R. antaRNA: ant colony-based RNA sequence design. Bioinformatics. 2015;31(19):3114–3121. https://doi.org/10.1093/bioinformatics/btv319.

[5] Runge F, Stoll D, Falkner S, Hutter F. Learning to Design RNA. In: Proceedings of the 7th International Conference on Learning Representations (ICLR 2019). New Orleans, LA, USA: OpenReview.net; 2019. Available from: https://openreview.net/forum?id=ByfyHh05tQ.

[6] Manzourolajdad A, Mohebbi M. Secondary-Structure-Informed RNA Inverse Design via Relational Graph Neural Networks. Non-Coding RNA. 2025;11(2):18. https://doi.org/10.3390/ncrna11020018.

[7] Wang J, Dokholyan NV. Modular Deep Learning for Direct RNA Sequence Design via Self-Contained RNA Units. bioRxiv. 2026;Preprint. Accessed: 2026-05-19. https://doi.org/10.64898/2026.04.16.719021.

[8] Huang H, Lin Z, He D, Hong L, Li Y. RiboDiffusion: tertiary structure-based RNA inverse folding with generative diffusion models. Bioinformatics. 2024;40(Supplement_1):i347–i356. Presented at ISMB 2024. Accessed: 2026-05-19. https://doi.org/10.1093/bioinformatics/btae259.

[9] Hu T, Cui Y, Luo B, Li K. RIDER: 3D RNA Inverse Design with Reinforcement Learning–Guided Diffusion. In: The Fourteenth International Conference on Learning Representations (ICLR); 2026. .

[10] Shen T, Hu Z, et al.: E2Efold-3D: End-to-End Deep Learning Method for Accurate De Novo RNA 3D Structure Prediction. arXiv preprint arXiv:2207.01586.

[11] Shen T, Hu Z, Sun S, Liu D, Wong F, Wang J, et al. Accurate RNA 3D structure prediction using a language model-based deep learning approach. Nature Methods. 2024;21:2287–2298. https://doi.org/10.1038/s41592-024-02487-0.

[12] Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630:493–500. https://doi.org/10.1038/s41586-024-07487-w.

[13] Peng C, Ni W, Liu Q, Hu G, Zheng W. A comprehensive benchmarking of the AlphaFold3 for predicting biomacromolecules and their interactions. Briefings in Bioinformatics. 2025;26(6):bbaf616. https://doi.org/10.1093/bib/bbaf616.

[14] Das R, Karanicolas J, Baker D. Atomic accuracy in predicting and designing noncanonical RNA structure. Nature Methods. 2010;7:291–294. https://doi.org/10.1038/nmeth.1433.