MORAD: REINFORCEMENT LEARNING WITH MULTI-OBJECTIVE REWARDS FOR RNA INVERSE DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce MORAD, an online reinforcement learning framework for backbone-conditioned RNA diffusion models that trains a single shared policy across diverse target backbones. Structure-conditioned RNA inverse design seeks nucleotide sequences compatible with a target three-dimensional backbone, yet existing reinforcement learning approaches either optimize a separate policy for each target backbone or rely on offline preference optimization, leaving online post-training of a reusable backbone-conditioned model largely unexplored. MORAD combines backbone-conditioned group sampling, group-relative reward normalization, and a forward-process diffusion objective. Beyond existing RL rewards that primarily target tertiary structural quality, MORAD jointly optimizes six measurements spanning tertiary geometry, base pairing, ensemble behavior, and sequence composition. To aggregate these heterogeneous objectives despite differences in scale and optimization direction, MORAD maps each measurement to a bounded desirability score and combines them with a weighted geometric mean. Compared with a tertiary-structure-only reward, this multi-objective formulation improves 10 of the 11 evaluation metrics that have a preferred direction. Applied to pre-trained RIDE, MORAD improves GDT-TS by 16.0% (from 0.3300 to 0.3827) and pairing MCC by 19.4% (from 0.6110 to 0.7295), while reducing normalized ensemble defect by 27.5% (from 0.3906 to 0.2830). On the test set, MORAD outperforms RiboDiffusion, gRNAde, and RDesign across tertiary structural accuracy, secondary-structure pairing, and thermodynamic quality, establishing a new state of the art in structure-conditioned RNA inverse design.

1 INTRODUCTION

RNA inverse design generates nucleotide sequences that are compatible with a specified structure. Three-dimensional backbone conditioning makes local geometry and long-range spatial relationships available to a sequence generator. This formulation enables the construction of candidate RNAs for subsequent folding and functional evaluation. It also defines a one-to-many learning problem: a target backbone can admit multiple compatible sequences. Recent structure-conditioned methods therefore model a conditional sequence distribution and evaluate both structural compatibility and sequence recovery (Joshi et al., 2025; Huang et al., 2024; Tan et al., 2024).

Geometric learning and diffusion modeling provide complementary ways to represent this distribution. gRNAde encodes RNA backbones with geometric graph neural networks and supports multiple conformations (Joshi et al., 2025). RDesign exploits hierarchical representation learning to use limited structural data efficiently (Tan et al., 2024). RiboDiffusion combines structural conditioning with generative denoising, while RIDE provides a structure-conditioned diffusion backbone within the RIDER framework (Huang et al., 2024; Hu et al., 2026). These models establish a strong starting distribution for design. Post-training can then further improve these models using properties measured on their newly generated sequences.

Reinforcement learning (RL) makes this feedback loop explicit. A design policy proposes candidates, a computational evaluator scores them, and the policy learns from the resulting rewards. LEARNA and Meta-LEARNA established this approach for secondary-structure-conditioned RNA design (Runge et al., 2019). More recent RNA methods investigate preference optimization, RL-guided diffusion, and step-wise latent optimization (Sun et al., 2025; Hu et al., 2026; Si et al., 2026). Protein

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.

design offers parallel evidence: EvoPlay combines self-play with sequence search, RL-DIF tunes structure-conditioned categorical diffusion, and ProteinZero updates inverse-folding models with online structural and stability feedback (Wang et al., 2023; Ektefaie et al., 2024; Wang et al., 2025b). Together, these developments motivate learning a reusable design policy from feedback collected across many RNA backbones.

Two coupled choices determine how such a policy learns. The first is the training interface. RNA feedback can involve a learned folding predictor, discrete base-pair comparisons, and thermodynamic software. An effective update rule must consume their scalar outputs while accommodating a multistep diffusion generator. Forward-process diffusion RL provides such an update rule through reward-weighted denoising objectives (Zheng et al., 2026). The second choice is the reward itself. Tertiary alignment, pairing compatibility, and ensemble behavior describe different aspects of a design, with different units and scales. Their aggregation determines which candidate-level improvements the policy prioritizes. A structured reward should preserve these distinctions while producing an effective within-target learning signal.

MORAD integrates these choices into a shared-policy RNA post-training workflow. For each target, an exponential-moving-average policy samples a group of sequences. Six measurements are mapped to bounded desirability scores, which a weighted geometric mean combines into a scalar reward. Group-relative normalization converts that reward into candidate weights for a DiffusionNFT-style update (Zheng et al., 2026; Shao et al., 2024). Training repeats across target backbones, and the resulting policy generates sequences directly from a new backbone at inference. We will publicly release the complete training code, data, and trained models.

Our contributions are:

1. **A reusable RNA online RL framework.** We formulate a shared-policy training procedure that connects structure-conditioned diffusion sampling, group-relative feedback, and forward-process optimization through modular interfaces.
2. **A coordinated multi-objective reward.** We specify six RNA design measurements, smooth desirability mappings, and weighted geometric aggregation, together with the data construction needed to support online scoring.
3. **An improved shared RNA design model.** We obtain MORAD by post-training RIDE and evaluate the same checkpoint across tertiary structure, secondary structure, and thermodynamics. On 153 evaluation targets, it reaches 0.3827 GDT-TS and 0.3400 TM-score. Held-out evaluations at 40 points during training track this improvement.

2 Related Work

Conditioned RNA design. Secondary-structure-conditioned design uses pairing constraints to guide sequence search or learning. LEARNA trains a design agent, and Meta-LEARNA transfers experience across target structures (Runge et al., 2019). Three-dimensional conditioning adds backbone geometry: gRNAde develops geometric, multi-state sequence modeling; RDesign learns hierarchical structural representations; and RiboDiffusion generates sequences through conditional denoising (Joshi et al., 2025; Tan et al., 2024; Huang et al., 2024). RIDER combines its RIDE backbone with RL-guided optimization, optimizing and reporting one target backbone at a time, so the resulting weights are target-specific rather than reusable; SOLD studies step-wise optimization in a latent diffusion model (Hu et al., 2026; Si et al., 2026). RNA3DB groups RNA structures into structurally dissimilar components, providing a basis for assessing transfer across RNA structures (Szikszai et al., 2024).

Reward-based learning for biomolecular sequences. Protein RL spans mutation search, language-model adaptation, and structure-conditioned generation. EvoPlay treats residue mutations as actions and couples a policy-value network with Monte Carlo tree search (Wang et al., 2023). KnowRLM incorporates biochemical knowledge into reinforced language models for directed evolution (Wang et al., 2024). RL-DIF optimizes a pretrained categorical diffusion model using structural consistency, with foldable diversity as an explicit evaluation goal (Ektefaie et al., 2024). ProteinZero uses online structural and stability proxies together with reference-policy and diversity regularization (Wang et al., 2025b). DRAKES develops reward optimization for discrete diffusion with applications to DNA

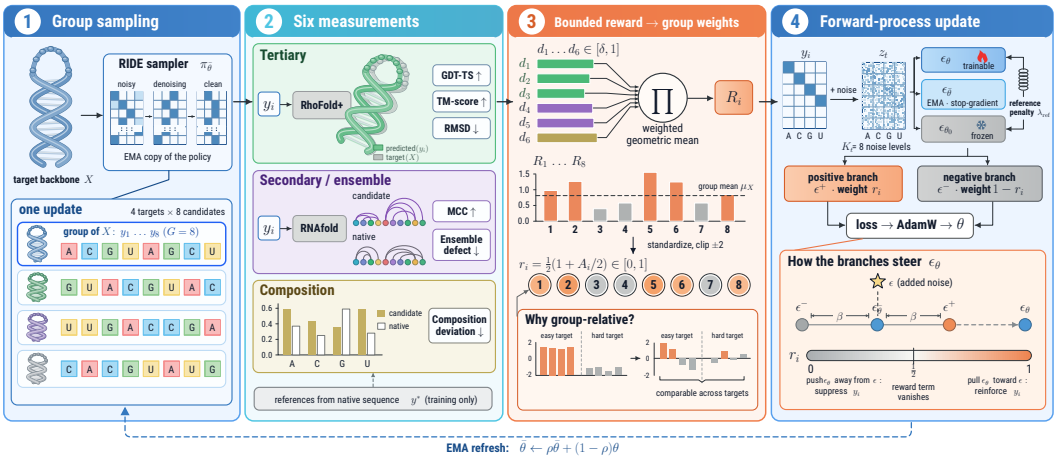

Figure 1: **MORAD training loop.** Backbone-conditioned sampling produces candidate sequences. Three evaluation channels provide the six-component reward and group-relative weights. Forward-noised candidates train the shared policy under a frozen-reference constraint; EMA feedback refreshes the sampler. The update follows the positive/negative objective in Eq. (3); the lower feedback path denotes EMA refresh.

and protein design (Wang et al., 2025a). For RNA, RiboPO constructs structure- and stability-aware preferences, complementing the online optimization routes explored by RIDER and SOLD (Sun et al., 2025; Hu et al., 2026; Si et al., 2026).

Several methods train generative models from reward feedback. DDPO views reverse diffusion as a sequential decision process (Black et al., 2024). DRAKES propagates reward information through relaxed discrete sampling, while DiffusionNFT learns through the forward process using positive and negative denoising branches (Wang et al., 2025a; Zheng et al., 2026). GRPO provides a group-relative baseline for reward normalization (Shao et al., 2024). These mechanisms address distinct parts of the learning loop: trajectory optimization, reward propagation, and candidate-level credit assignment.

Multi-objective measurement and evaluation. Desirability functions map responses with different units onto a common scale; geometric aggregation is a classical construction for coordinating multiple responses (Derringer & Suich, 1980). RNA evaluation combines learned structural predictors such as RhoFold+, secondary-structure predictors such as EternaFold, and energy-based calculations from ViennaRNA (Shen et al., 2024; Wayment-Steele et al., 2022; Lorenz et al., 2011). US-align supports structural comparisons of nucleic acids (Zhang et al., 2022). RiboPO organizes evaluation across sequence, secondary structure, tertiary structure, and thermodynamics (Sun et al., 2025). Uncertainty estimates also complement aggregate RL scores (Agarwal et al., 2021).

3 Method

3.1 Online training algorithm and optimization objective

Let X denote an RNA backbone with L valid nucleotide positions and let $y \in \{A, C, G, U\}^L$ denote a designed sequence. A pretrained conditional diffusion model initializes three parameter sets: the trainable policy θ, the moving-average sampling policy $\bar{\theta}$, and the frozen reference θ_0. The online objective is to increase $\mathbb{E}_{X\sim\mathcal{D}_{\text{train}},\, y\sim\pi_\theta(\cdot|X)}[R(X,y)]$ while keeping a reference-based constraint on denoising predictions. Figure 1 summarizes the training loop.

Target-relative feedback. For each sampled target, $\pi_{\bar{\theta}}$ generates G candidates. The scorer returns valid rewards $\{R_i\}_{i\in\mathcal{V}_X}$. We use population statistics over the valid group and define

$$\mu_X = \frac{1}{|\mathcal{V}_X|}\sum_{i\in\mathcal{V}_X} R_i, \quad s_X^2 = \frac{1}{|\mathcal{V}_X|}\sum_{i\in\mathcal{V}_X}(R_i-\mu_X)^2, \quad A_i = \text{clip}\left(\frac{R_i-\mu_X}{s_X+\varepsilon_A}, -z_c, z_c\right). \tag{1}$$

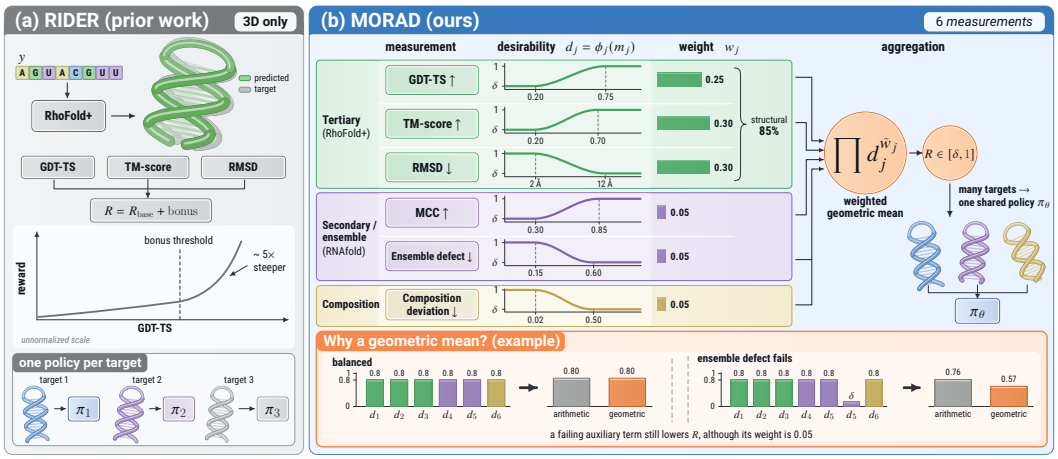

Figure 2: **RIDER and MORAD reward construction.** RIDER scores tertiary structural consistency (Hu et al., 2026). MORAD combines three tertiary measurements, pairing agreement, normalized ensemble defect (ED), and composition deviation. Each desirability score lies in $[\delta, 1]$; ϕ_j denotes the map in Eq. (4), and $\widetilde{w}_j$ is the normalized weight in Eq. (5). The panels compare reward definitions.

The candidate weight is $r_i = \frac{1}{2}(1 + A_i/z_c)$, with $z_c = 2$ and $\varepsilon_A = 10^{-8}$. Group normalization compares sequences under the same structural condition and places feedback on a common scale across targets. Groups require at least two valid rewards and positive reward variance to contribute a policy gradient. Candidates are drawn at sampling temperature $\tau = 1$; a group returning fewer than two valid rewards or fewer than two distinct sequences is resampled once with a new seed and the temperature raised by 0.1, and it is dropped if the second attempt also fails. Rewards, group statistics, and r_i are treated as constants during optimization.

Forward-process policy update. The one-hot sequence $x_0 \in \{0,1\}^{L\times 4}$ is corrupted as $z_t = \alpha_t x_0 + \sigma_t \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and $\alpha_t^2 + \sigma_t^2 = 1$ (Ho et al., 2020). We sample K_t stratified log-signal-to-noise levels per candidate. Let ϵ_θ, $\epsilon_{\bar{\theta}}$, and ϵ_{θ_0} denote the corresponding conditional noise predictions at (z_t, t, X). DiffusionNFT is formulated with flow-matching velocity prediction (Zheng et al., 2026); we transfer its positive/negative construction to the ϵ-prediction space of RIDE and form

$$\epsilon^+ = (1-\beta)\epsilon_{\bar{\theta}} + \beta\epsilon_\theta, \qquad \epsilon^- = (1+\beta)\epsilon_{\bar{\theta}} - \beta\epsilon_\theta, \tag{2}$$

with stop-gradient applied to $\epsilon_{\bar{\theta}}$. The scale $\beta \in (0, 1]$ sets how far the two branches are displaced from the sampling policy. The objective is

$$\mathcal{L}(\theta) = \mathbb{E}\left[\frac{r_i \left\|\epsilon^+ - \epsilon\right\|_{\text{av}}^2 + (1-r_i)\left\|\epsilon^- - \epsilon\right\|_{\text{av}}^2}{\beta} + \lambda_{\text{ref}} \left\|\epsilon_\theta - \text{sg}(\epsilon_{\theta_0})\right\|_{\text{av}}^2\right]. \tag{3}$$

Here $\|\cdot\|_{\text{av}}^2$ averages squared residuals across sequence positions and channels; the expectation averages valid candidates, targets, sampled noise, and time levels. We use uniform weighting in noise-prediction space. The reference term penalizes changes to the pretrained denoiser, and the positive/negative branches turn relative reward into an update direction. Because of the $1/\beta$ factor, the cross term that carries the reward does not depend on β, so β scales only the quadratic displacement penalty; we use $\beta = 0.1$ and $\lambda_{\text{ref}} = 0.02$ (Appendix D). Appendix A derives this relation and the equivalent clean-sequence weighting.

We use AdamW (Loshchilov & Hutter, 2019). After each optimizer step, $\bar{\theta} \leftarrow \rho_u \bar{\theta} + (1-\rho_u)\theta$, where $\rho_u = \min(0.999, (1+u)/(10+u))$ at update u. The frozen reference remains fixed. Algorithm 1 lists each step of the training loop. It uses $\ell_{\theta,X,i,k}$ for the integrand of Eq. (3), p_{strat} for stratified time sampling, and clip_1 for unit-norm gradient clipping. GroupStats returns the valid indices, reward mean, and population standard deviation. The procedure requires only a conditional sampler, a noise-prediction adapter, and scalar evaluators. The experiments instantiate these components with RIDE; the trained policy shares parameters across all target backbones.

Algorithm 1 MORAD online post-training

Require: Training data $\mathcal{D}_{\text{train}}$; pretrained parameters θ_0; group size G; sampling temperature τ; noise draws K_t; updates U; loss coefficients $\beta, \lambda_{\text{ref}}$

Initialize policies: $\theta \leftarrow \theta_0$, $\bar{\theta} \leftarrow \theta_0$; keep θ_0 frozen.
for $u = 1, \ldots, U$ **do**
  Sample targets: $\mathcal{B}_u \sim \mathcal{D}_{\text{train}}$.
  Generate candidates: $y_{X,1:G} \sim \pi_{\bar{\theta}}(\cdot \mid X)$ at temperature τ for each $X \in \mathcal{B}_u$.
  Score designs: $R_{X,i} \leftarrow R(X, y_{X,i})$ using Eq. (5).
  Resample degenerate groups once: repeat the two steps above at temperature $\tau + 0.1$ for any X with fewer than two valid rewards or fewer than two distinct sequences.
  Summarize groups: $(\mathcal{V}_X, \mu_X, s_X) \leftarrow \text{GroupStats}(R_{X,1:G})$.
  Retain informative groups: $\mathcal{B}_u^+ \leftarrow \{X \in \mathcal{B}_u : |\mathcal{V}_X| \geq 2,\ s_X > 0\}$.
  for all valid candidates $X \in \mathcal{B}_u^+,\ i \in \mathcal{V}_X$ **do**
    Normalize feedback: compute $A_{X,i}$ by Eq. (1).
    Detach weights: $r_{X,i} \leftarrow \text{sg}\big(\frac{1}{2}(1 + A_{X,i}/z_c)\big)$.
    Draw noise and times: $(t_{ik}, \epsilon_{ik}) \sim p_{\text{strat}}(t)\,\mathcal{N}(0, I)$, $k = 1{:}K_t$.
    Perturb sequences: $z_{ik} \leftarrow \alpha_{t_{ik}} \text{onehot}(y_{X,i}) + \sigma_{t_{ik}} \epsilon_{ik}$.
  end for
  Aggregate the two-branch loss from Eq. (3) over valid candidates:

$$\mathcal{L}_u \leftarrow \frac{1}{|\mathcal{B}_u|} \sum_{X \in \mathcal{B}_u^+} \frac{1}{|\mathcal{V}_X| K_t} \sum_{i \in \mathcal{V}_X} \sum_{k=1}^{K_t} \ell_{\theta,X,i,k}.$$

  Update the policy: $\theta \leftarrow \text{AdamW}(\theta, \text{clip}_1(\nabla_\theta \mathcal{L}_u))$.
  Refresh the EMA sampler: $\bar{\theta} \leftarrow \rho_u \bar{\theta} + (1 - \rho_u)\theta$.
end for
return shared design policy $\pi_\theta(\cdot \mid X)$.

3.2 Reward construction

The reward combines six quantities with explicit roles. GDT-TS, TM-score, and RMSD assess the RhoFold+-predicted structure against the target backbone (Shen et al., 2024). Secondary-structure MCC measures agreement between the RNAfold structures of a candidate and its native reference (Lorenz et al., 2011; Matthews, 1975). Normalized ensemble defect measures expected nucleotide-level disagreement with that reference pairing under the candidate's thermodynamic ensemble (Zadeh et al., 2011). A composition-deviation term compares candidate and native A/C/G/U frequencies. Appendix D.2 gives the measurement formulas and evaluation-tool configuration.

For metric m_j, anchors a_j and b_j identify low and high desirability. We apply a clipped raised-cosine map,

$$u_j = \text{clip}\left(\frac{m_j - a_j}{b_j - a_j}, 0, 1\right), \qquad d_j = \max\left(\delta, \frac{1 - \cos(\pi u_j)}{2}\right), \quad \delta = 10^{-3}. \tag{4}$$

Lower-is-better measurements use $b_j < a_j$. The anchors align units and specify the region where changes in a measurement affect the reward. Saturation limits the influence of extreme values, and the floor keeps logarithms finite. The scalar reward is

$$R(X, y) = \exp\left(\sum_{j \in \mathcal{J}(X,y)} \widetilde{w}_j \log d_j\right), \qquad \widetilde{w}_j = \frac{w_j}{\sum_{k \in \mathcal{J}(X,y)} w_k}, \tag{5}$$

where $\mathcal{J}$ contains available measurements. The weights for GDT-TS, TM-score, RMSD, MCC, ensemble defect, and composition are $(0.25, 0.30, 0.30, 0.05, 0.05, 0.05)$. Valid scoring requires the three structural terms; an unavailable auxiliary measurement is dropped from $\mathcal{J}$, and Eq. (5) renormalizes over the remaining active weights.

Geometric aggregation favors candidates that score reasonably on every measurement over candidates that excel on only one (Derringer & Suich, 1980). Above the desirability floor, $\partial \log R / \partial d_j = \widetilde{w}_j / d_j$, so at a fixed weight the reward responds most strongly to the lowest scores. This property motivates the geometric form; it reaches the optimizer through the candidate weights r_i. Structural terms carry 85% of the nominal weight, while the three auxiliary terms still directly affect how candidates are ranked. With Eq. (1), normalization has two stages: desirability scores align the units and directions

Table 1: **Tertiary-structure evaluation.** First sampled candidate (candidate 0) for each of the 153 evaluation targets; C1′ TM has 126 valid targets. MORAD is the checkpoint at update 390. Alignment scores and additional diagnostics are shown together; RMSD is in Å. Bold marks the best mean. Metric definitions are in Appendix D.2.

Model	GDT-TS ↑	TM(C4′) ↑	RMSD (Å) ↓	TM(C1′) ↑	Success ↑	Recovery ↑	lDDT (C4′) ↑	pLDDT ↑	Coarse clash ↓
RiboDiffusion	0.3132	0.2839	10.9215	0.3613	0.2549	**0.5165**	0.6174	**0.6937**	73.5383
gRNAde ($T = 0.1$)	0.3591	0.3179	8.8374	0.3642	0.2484	0.5115	0.6607	0.6482	23.7261
RDesign	0.3291	0.2902	10.3441	0.3392	0.2549	0.4720	0.6368	0.6202	40.4457
RIDE	0.3300	0.2841	9.7199	0.3538	0.2484	0.5039	0.6622	0.6671	22.7503
MORAD	**0.3827**	**0.3400**	**7.9196**	**0.3855**	**0.3007**	0.5056	**0.6774**	0.6726	**11.9438**

of the six measurements, and group-relative standardization removes difficulty differences between targets. Appendix B lists all anchors and the structural-only control reward.

3.3 Data construction

Each record stores a backbone X, its native sequence $y^\star$, and a coordinate-validity mask. The backbone conditions generation, and the native sequence provides pairing and composition references during training; at inference the policy conditions on the target backbone alone. The training set contains 527 targets and the evaluation set 153. The evaluation set is built separately from the training set and comprises 71 entries from the RNA3DB test split and 82 independently curated RCSB entries spanning 15–186 nucleotides (Szikszai et al., 2024; Berman et al., 2000). No evaluation target shares a PDB entry with a training target, and no evaluation sequence exactly matches a training sequence after normalization; each evaluation target's maximum sequence identity to the training set has a median of 0.62, and Appendix C reports its full distribution. All 153 evaluation targets are used. Appendix C details preprocessing and screening.

4 Experiments

4.1 Experimental setup

We initialize from pretrained RIDE and train for 390 updates over three epochs on four GPUs, using four target groups of eight candidates per update. The reported model is the final checkpoint, update 390; the held-out evaluations taken every ten updates document training and are not used to select a checkpoint. The 10.2M-parameter model uses an initial learning rate of 2×10^{-4} with cosine decay to 0.3 times the initial rate. RhoFold+ scores tertiary structure (Shen et al., 2024); Appendix D gives the configuration.

Tables 1 and 2 compare MORAD with RiboDiffusion, gRNAde ($T = 0.1$), RDesign, and pretrained RIDE using the first sampled candidate (candidate 0) for each of the 153 evaluation targets (Huang et al., 2024; Joshi et al., 2025; Tan et al., 2024; Hu et al., 2026); the index is fixed for every model, with no best-of-eight selection. C1′ TM has 126 valid targets; pairing counts vary by metric (Appendix D.1.1). Diversity uses all eight candidates per target (Appendix D.1.2); Table 3 reports a separate thermodynamic evaluation that also includes gRNAde ($T = 0.8$). The checkpoint comparison (Appendix E) and the training curves (Figure 3) come from separate evaluation runs.

4.2 Comparison with existing RNA design models

Across Tables 1–3, MORAD attains the best mean on 14 of the 16 metrics with a preferred direction; the two exceptions, recovery and pLDDT, are led by RiboDiffusion.

Tertiary structure. MORAD improves GDT-TS by 16.0% over RIDE (from 0.3300 to 0.3827) and TM-score by 19.7% (from 0.2841 to 0.3400), while reducing RMSD from 9.7199 to 7.9196 Å and raising C1′ TM from 0.3538 to 0.3855 (Table 1). It also leads success rate, C4′ lDDT and coarse

Table 2: **Candidate-0 secondary-structure agreement.** EternaFold predictions are compared with base pairs extracted from the native 3D structures (Wayment-Steele et al., 2022). Bold marks the best mean. Valid counts for each metric and model are in Appendix D.1.1; thermodynamic defect is reported separately in Table 3.

Model	Precision ↑	Recall ↑	F1 ↑	MCC ↑
RiboDiffusion	0.3563	0.3074	0.3212	0.3267
gRNAde ($T = 0.1$)	0.7276	0.6099	0.6482	0.6734
RDesign	0.7084	0.6110	0.6462	0.6566
RIDE	0.6472	0.5717	0.5939	0.6110
MORAD	**0.7824**	**0.6816**	**0.7226**	**0.7295**

clash, while RiboDiffusion leads recovery and pLDDT. These metrics capture different properties: structural agreement, predictor confidence, and sequence recovery.

Secondary-structure agreement. MORAD improves pairing MCC by 19.4% over RIDE (from 0.6110 to 0.7295) and leads all four metrics: precision 0.7824, recall 0.6816, F1 0.7226, and MCC 0.7295, versus RIDE's 0.6472, 0.5717, 0.5939, and 0.6110 (Table 2). Neither EternaFold nor these structure-derived base pairs enter the reward, which scores pairing with RNAfold (Appendix B); these pairing gains are therefore measured with tools outside the reward.

Thermodynamic ensemble profile. Table 3 reports the thermodynamic evaluation on all 153 targets. MORAD leads all three metrics with a preferred direction: NED drops from RIDE's 0.3906 to 0.2830, a 27.5% reduction in normalized target-structure defect, PE from 0.3059 to 0.2102, and MFEfreq rises from 0.3350 to 0.4402, so the ensembles concentrate more tightly on the target (McCaskill, 1990; Lorenz et al., 2011; Zadeh et al., 2011). Free energies are shown for context and are not ranked: MORAD's MFE/EFE, $-30.149/-30.937$ kcal/mol, are the most stable here yet the furthest from the native -21.261. Appendix D.3 reports the remaining metrics: absolute ED, ensemble diversity, melting temperature, and GC content.

4.3 Reward Design

Existing RL rewards for RNA design score tertiary structure alone; RIDER, for example, combines structural scores with threshold bonuses (Hu et al., 2026). MORAD instead combines six bounded desirability scores, so that pairing agreement, ensemble behavior, and composition share the learning signal with tertiary geometry (Figure 2). To isolate this choice, we post-train two policies from the same pretrained RIDE using the same data, optimizer, and schedule, and change only the reward: a structural-only reward of the kind used by existing work, and the MORAD reward. Table 4 reports both policies alongside their pretrained starting point.

The MORAD reward leads on ten of the eleven metrics with a preferred direction: tertiary geometry (GDT-TS 0.3627 to 0.3694; TM-score 0.3194 to 0.3358; RMSD 8.369 to 8.207 Å; C1′ TM 0.3818 to 0.3919), pairing agreement (MCC 0.6612 to 0.6978; F1 0.6568 to 0.6901), ensemble defect

Table 3: **Thermodynamic evaluation on 153 targets.** MFE and EFE are minimum and ensemble free energies (kcal/mol); NED is normalized ensemble defect; PE is positional entropy; MFEfreq is the probability of the minimum-energy structure (Lorenz et al., 2011; Zadeh et al., 2011). MORAD is update 390. Bold marks the best mean; ∼ marks metrics without a preferred direction, shown for context. Appendix D.3 reports all nine metrics and the native reference.

Model	MFE ∼	EFE ∼	NED ↓	PE ↓	MFEfreq ↑
RiboDiffusion	-21.716	-22.847	0.6056	0.4284	0.2980
gRNAde ($T = 0.1$)	-28.092	-28.918	0.3458	0.2652	0.4289
gRNAde ($T = 0.8$)	-25.186	-26.212	0.3830	0.3260	0.3266
RDesign	-28.863	-29.708	0.3349	0.2775	0.3961
RIDE	-23.614	-24.643	0.3906	0.3059	0.3350
MORAD	-30.149	-30.937	**0.2830**	**0.2102**	**0.4402**

Table 4: **Reward comparison under an identical training setup.** Two policies are post-trained from the same pretrained RIDE with the same data, optimizer, and schedule, and differ only in the reward; each is shown at the last checkpoint of its run. The structural-only reward matches the scope of existing RNA design RL rewards and is defined in Appendix B. These results come from the controlled experiment's own evaluation run, so some entries differ from Tables 1–3. Bold marks the best of the three for metrics with a preferred direction; MFE, GC fraction, and within-target Hamming diversity are shown for context.

Measurement	RIDE	+ Structural reward	+ MORAD reward
Tertiary structure			
GDT-TS ↑	0.3300	0.3627	**0.3694**
TM-score ↑	0.2841	0.3194	**0.3358**
RMSD (Å) ↓	9.720	8.369	**8.207**
C1′ TM ↑	0.3538	0.3818	**0.3919**
Secondary structure and recovery			
MCC ↑	0.6208	0.6612	**0.6978**
F1 ↑	0.6103	0.6568	**0.6901**
ED (nt) ↓	29.08	26.85	**24.27**
ED/nt ↓	0.3872	0.3569	**0.3256**
Recovery ↑	0.5033	0.5097	**0.5124**
Equilibrium ensemble			
H_{pair} ↓	0.3249	0.2851	**0.2800**
p_{MFE} ↑	0.3281	**0.3880**	0.3847
MFE (kcal/mol)	-23.77	-29.81	-26.49
GC fraction	0.5304	0.5912	0.4970
Candidate diversity			
Hamming	0.2971	0.2224	0.2607

(ED/nt 0.3569 to 0.3256), recovery (0.5097 to 0.5124), and pairing entropy (0.2851 to 0.2800). MFE-structure frequency is the one measurement where the structural-only reward is higher, 0.3880 against 0.3847, and it comes with a shift in composition: that policy reaches GC fraction 0.5912 and MFE -29.81 kcal/mol, while the MORAD reward stays at 0.4970 and -26.49 kcal/mol, both closer to the native values of 0.5207 and -21.26 kcal/mol. Since only the reward differs, these differences follow from the reward design (Appendix B).

Effect of the auxiliary terms. Although the auxiliary terms carry only 15% of the nominal weight, the MORAD reward also improves the shared tertiary scores: geometric aggregation lowers the reward of candidates that align well but pair poorly.

4.4 Held-out performance during training

We evaluate the held-out targets every tenth update, giving 40 measurements from update 0 to 390 (Figure 3). GDT-TS increases from 0.3365 to 0.3711 and C4′ TM from 0.2912 to 0.3296: relative gains of 10.3% and 13.2%. These values come from the held-out evaluations run during training; Table 1 is a separate, final evaluation of update 390.

Candidate variation. During training, within-target sequence diversity decreases from 0.2023 to 0.1616. In the final eight-candidate evaluation, MORAD has Hamming and 3-mer diversity of 0.2344 and 0.3673, versus RIDE's 0.2971 and 0.5270; Appendix D.1.2 gives all models. This quality–diversity tradeoff is well-documented in language-model RL and protein inverse folding (Kirk et al., 2024; Yue et al., 2025; Ektefaie et al., 2024; Wang et al., 2025b), motivating joint reporting of both dimensions.

5 Limitations

The experiments evaluate MORAD on a single 10.2M-parameter RIDE backbone, demonstrating the framework on a compact model; the same modular interfaces accommodate larger structure-conditioned backbones. Reward scoring and tertiary evaluation both use the RhoFold+ predictor,

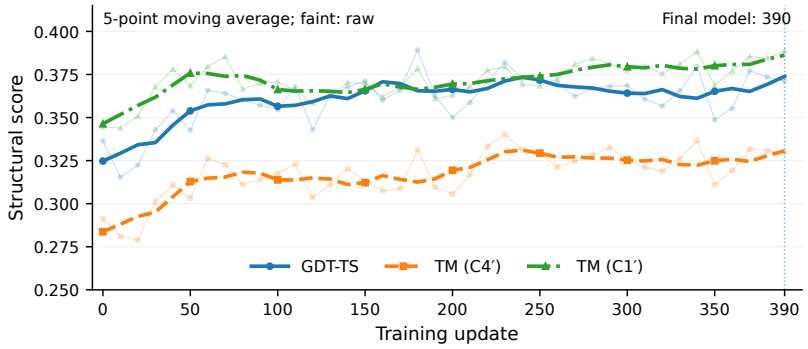

Figure 3: **Structural-score trends over all 40 evaluations.** Thick lines show a centered five-point moving average; faint traces show the raw held-out scores recorded during training, with one sampled candidate per target. Near the two ends, the average uses only the available points. Table 10 reports the evaluation of the last checkpoint, update 390. Appendix E.1 defines the smoothing.

consistent with current RNA design RL practice (Hu et al., 2026; Shen et al., 2024); experimental structural validation remains a separate step beyond the scope of this computational study. The reward also covers equilibrium folding properties, leaving kinetics and function-specific objectives as future additions to the reward.

6 Conclusion

MORAD establishes a shared-policy online post-training workflow for RNA design, integrating target-relative feedback, six-component rewards, and forward-process diffusion optimization. The last checkpoint of the completed run (update 390) reaches mean scores of 0.3827 GDT-TS and 0.3400 TM-score in the final evaluation. Together with the pairing and ensemble-defect results and the full training trajectory, these results document the quality gains and multi-objective tradeoffs of online post-training.

AI use statement

Language-model assistance supported manuscript revision and translation, consistency checking, and code for tables, training plots, and document compilation. The two method diagrams (Figures 1 and 2) were drawn by the authors. Empirical measurements come from the authors' experiment records; no new measurements or training runs were produced during manuscript preparation. The authors are responsible for the scientific claims, evidence, references, and final submission.

Ethics statement

This is a computational study. It uses publicly available RNA structures from RNA3DB and the PDB and involves no human subjects, no animal experiments, and no new laboratory measurements (Szikszai et al., 2024; Berman et al., 2000). Designed sequences are computational candidates scored by folding and thermodynamic software; none were synthesized or assayed, and we make no functional, therapeutic, or biosafety claims for them.

Reproducibility statement

Algorithm 1 and Eqs. (1)–(5) specify the training procedure and the reward. Appendix B lists every anchor and weight, Appendix D the run configuration, and Appendix D.1 the aggregation rules and valid-observation counts; Appendix C documents the data splits and screening. We will publicly release the complete training code, data, and trained models, together with the trajectory, checkpoint, and measurement records, the table- and figure-generation scripts, and the automated consistency checks.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320, 2021. URL https://arxiv.org/abs/2108.13264.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.13301.

COLA Laboratory. RIDER reinforcement learning implementation: trainer_rl.py. Source code, revision 799f9a0, 2026. URL https://github.com/COLA-Laboratory/RIDER/blob/799f9a0f6587b725c6319dcc24e423e65bcc3682/trainer_rl.py.

George Derringer and Ronald Suich. Simultaneous optimization of several response variables. *Journal of Quality Technology*, 12(4):214–219, 1980. doi: 10.1080/00224065.1980.11980968.

Yasha Ektefaie, Olivia Viessmann, Siddharth Narayanan, Drew Dresser, J. Mark Kim, and Armen Mkrtchyan. Reinforcement learning on structure-conditioned categorical diffusion for protein inverse folding. *arXiv preprint arXiv:2410.17173*, 2024. URL https://arxiv.org/abs/2410.17173.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2006.11239.

Tianmeng Hu, Yongzheng Cui, Biao Luo, and Ke Li. RIDER: 3D RNA inverse design with reinforcement learning-guided diffusion. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.16548.

Han Huang, Ziqian Lin, Dongchen He, Liang Hong, and Yu Li. RiboDiffusion: Tertiary structure-based RNA inverse folding with generative diffusion models. *Bioinformatics*, 40(Supplement 1): i347–i356, 2024. doi: 10.1093/bioinformatics/btae259.

Chaitanya K. Joshi, Arian R. Jamasb, Ramon Viñas, Charles Harris, Simon V. Mathis, Alex Morehead, Rishabh Anand, and Pietro Liò. gRNAde: Geometric deep learning for 3D RNA inverse design. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2305.14749.

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.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06452.

Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6:26, 2011. doi: 10.1186/1748-7188-6-26.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451, 1975. doi: 10.1016/0005-2795(75)90109-9.

J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6–7):1105–1119, 1990. doi: 10.1002/bip.360290621.

Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3): 443–453, 1970. doi: 10.1016/0022-2836(70)90057-4.

Frederic Runge, Danny Stoll, Stefan Falkner, and Frank Hutter. Learning to design RNA. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1812.11951.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

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

Qi Si, Xuyang Liu, Penglei Wang, Xin Guo, Yuan Qi, and Yuan Cheng. Structure-based RNA design by step-wise optimization of latent diffusion model. In *AAAI Conference on Artificial Intelligence*, 2026. URL https://arxiv.org/abs/2601.19232.

Minghao Sun, Hanqun Cao, Zhou Zhang, Chen Wei, Liang Wang, Tianrui Jia, Zhiyuan Liu, Tianfan Fu, Xiangru Tang, Yejin Choi, Pheng-Ann Heng, Fang Wu, and Yang Zhang. RiboPO: Preference optimization for structure- and stability-aware RNA design. *arXiv preprint arXiv:2510.21161*, 2025. URL https://arxiv.org/abs/2510.21161.

Marcell Szikszai, Marcin Magnus, Siddhant Sanghi, Sachin Kadyan, Nazim Bouatta, and Elena Rivas. RNA3DB: A structurally-dissimilar dataset split for training and benchmarking deep learning models for RNA structure prediction. *Journal of Molecular Biology*, 436(17):168552, 2024. doi: 10.1016/j.jmb.2024.168552.

Cheng Tan, Yijie Zhang, Zhangyang Gao, Bozhen Hu, Siyuan Li, Zicheng Liu, and Stan Z. Li. RDesign: Hierarchical data-efficient representation learning for tertiary structure-based RNA design. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2ad9a1a6ffac3dd72cc1df96019eca01-Abstract-Conference.html.

Chenyu Wang, Masatoshi Uehara, Yichun He, Amy Wang, Tommaso Biancalani, Avantika Lal, Tommi Jaakkola, Sergey Levine, Hanchen Wang, and Aviv Regev. Fine-tuning discrete diffusion models via reward optimization with applications to DNA and protein design. In *International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2410.13643.

Yi Wang, Hui Tang, Lichao Huang, Lulu Pan, Lixiang Yang, Huanming Yang, Feng Mu, and Meng Yang. Self-play reinforcement learning guides protein engineering. *Nature Machine Intelligence*, 5(8):845–860, 2023. doi: 10.1038/s42256-023-00691-9.

Yuhao Wang, Qiang Zhang, Ming Qin, Xiang Zhuang, Xiaotong Li, Zhichen Gong, Zeyuan Wang, Yu Zhao, Jianhua Yao, Keyan Ding, and Huajun Chen. Knowledge-aware reinforced language models for protein directed evolution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52260–52273, 2024. URL https://proceedings.mlr.press/v235/wang24cq.html.

Ziwen Wang, Jiajun Fan, Ruihan Guo, Thao Nguyen, Heng Ji, and Ge Liu. ProteinZero: Self-improving protein generation via online reinforcement learning. *arXiv preprint arXiv:2506.07459*, 2025b. URL https://arxiv.org/abs/2506.07459.

Hannah K. Wayment-Steele, Wipapat Kladwang, Alexandra I. Strom, Jeehyung Lee, Adrien Treuille, Alex Becka, Eterna Participants, and Rhiju Das. RNA secondary structure packages evaluated and improved by high-throughput experiments. *Nature Methods*, 19:1234–1242, 2022. doi: 10.1038/s41592-022-01605-0.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025. URL https://arxiv.org/abs/2504.13837.

Joseph N. Zadeh, Brian R. Wolfe, and Niles A. Pierce. Nucleic acid sequence design via efficient ensemble defect optimization. *Journal of Computational Chemistry*, 32(3):439–452, 2011. doi: 10.1002/jcc.21633.

Adam Zemla. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Research*, 31(13):3370–3374, 2003. doi: 10.1093/nar/gkg571.

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

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004. doi: 10.1002/prot.20264.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.16117.

A FORWARD-PROCESS OBJECTIVE

For one candidate and sampled noise level, write $b = \epsilon_{\bar{\theta}} - \epsilon$ and $\Delta = \epsilon_\theta - \epsilon_{\bar{\theta}}$. The two branches in Eq. (2) give

$$\ell_{\text{policy}} = \frac{r\,\|b + \beta\Delta\|_{\text{av}}^2 + (1-r)\,\|b - \beta\Delta\|_{\text{av}}^2}{\beta} \tag{6}$$

$$= \frac{\|b\|_{\text{av}}^2}{\beta} + 2(2r-1)\langle b, \Delta\rangle_{\text{av}} + \beta\,\|\Delta\|_{\text{av}}^2\,. \tag{7}$$

Here the average divides by the number D of scalar prediction coordinates. Treating the sampling-policy prediction $\epsilon_{\bar{\theta}}$ and the weight r as constants,

$$\nabla_{\epsilon_\theta}\ell_{\text{policy}} = \frac{2}{D}\big[(2r-1)b + \beta\Delta\big]. \tag{8}$$

At $\Delta = 0$, gradient descent moves the prediction toward the sampled noise for $r > 1/2$ and in the opposite direction for $r < 1/2$. At $r = 1/2$, only the displacement penalty remains. The reference term in Eq. (3) further penalizes deviation from the pretrained predictor. This expansion shows how the reward enters the DiffusionNFT-style update (Zheng et al., 2026).

Parameterization and noise sampling. For a clean-sequence predictor $\widehat{x}_{0,\theta}$, the Gaussian parameterization (Ho et al., 2020) gives

$$\epsilon_\theta = \frac{z_t - \alpha_t\widehat{x}_{0,\theta}}{\sigma_t}, \qquad \|\widehat{x}_{0,\theta} - x_0\|_{\text{av}}^2 = \frac{\sigma_t^2}{\alpha_t^2}\,\|\epsilon_\theta - \epsilon\|_{\text{av}}^2\,. \tag{9}$$

The implementation uses uniform noise-space weighting with no residual normalization. Each candidate receives $K_t = 8$ stratified log-SNR draws from $[-10, 10]$. For log-SNR ℓ, it uses $t = \text{sigmoid}(-\ell)$, $\alpha_t = \sqrt{1-t}$, and $\sigma_t = \sqrt{t}$. Candidate one-hot sequences and feedback weights are treated as constants; gradients flow only through the current denoiser.

B REWARD PARAMETERS AND REFERENCE SCHEMES

Table 5: **Six-component reward parameters.** Anchors a_j, b_j identify low and high desirability. Equation (4) uses floor 10^{-3}. All three structural measurements must be finite; available auxiliary terms determine the active weights.

Measurement	a_j	b_j	w_j
GDT-TS	0.20	0.75	0.25
TM-score	0.20	0.70	0.30
RMSD (Å)	12.0	2.0	0.30
Secondary-structure MCC	0.30	0.85	0.05
Ensemble defect / nucleotide	0.60	0.15	0.05
Composition deviation	0.50	0.02	0.05

Auxiliary references. During training, RNAfold predicts both candidate and native secondary structures (Lorenz et al., 2011). Their base-pair sets define MCC (Matthews, 1975); the native RNAfold structure also defines the ensemble-defect reference (Zadeh et al., 2011). For candidate and native base frequencies p_b, q_b, the implemented composition term is

$$C(y, y^\star) = \max\left(0,\ \sum_{b:p_b>0} p_b \log\frac{p_b}{\max(q_b, 10^{-3})}\right). \tag{10}$$

Logarithms are natural. An unavailable auxiliary measurement is removed from the active set, followed by the weight normalization in Eq. (5).

Structural-only control reward. The control scores tertiary geometry alone, matching the scope of existing RNA design RL rewards (Hu et al., 2026): $R_{\text{ctrl}} = 0.45\phi_{.15,.70}(g) + 0.35\phi_{.20,.70}(t) + 0.20\phi_{12,1.5}(d)$, where g, t, d are GDT-TS, TM-score, and RMSD, and $\phi_{a,b}(m) = \frac{1}{2}[1 - \cos(\pi \,\text{clip}((m-a)/(b-a), 0, 1))]$. The MORAD reward of Eq. (5) adds pairing agreement, ensemble defect, and composition on the same desirability scale. The two policies in Table 4 differ in this reward alone; the Hamming diversity reported there averages pairwise normalized sequence distances within a target.

Published RIDER reward. RIDER defines four structural base rewards $R_{\text{base}} \in \{(5g)^2, (5t)^2, -(0.5d)^2, (5g)^2 - (0.5d)^2\}$ with the published bonus (Hu et al., 2026)

$$R_{\text{RIDER}} = R_{\text{base}} + \begin{cases} 100(g - 0.5), & g > 0.5, \\ 20(2 - d), & g \le 0.5,\ d < 2, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

The released RIDER code uses the combined base reward with thresholds 0.45 and 3 Å (COLA Laboratory, 2026). This reward is defined and optimized for a single target backbone, whereas a shared policy requires a reward that is comparable across targets; the control above keeps the structural-only scope while meeting that requirement. Figure 2 compares the two designs.

C Data construction and split characterization

Records and screening. Each record contains a native sequence, its source PDB chain, backbone coordinates, and a coordinate-validity mask. Normalization maps letters to uppercase and T to U; scoring uses A/C/G/U. Residues with 3D coordinates are matched to sequence positions. Screening folds the native sequence with RhoFold+ and keeps targets whose prediction reaches C4′ GDT-TS of at least 0.30 under centered, index-preserving Kabsch alignment (Kabsch, 1976).

Fixed splits. The training set has 527 targets of length 27–258 nt (median 75); the evaluation set has 153 targets of length 15–186 nt (median 65). Evaluation sources comprise 71 entries from the RNA3DB test split and 82 entries curated from RCSB (Szikszai et al., 2024; Berman et al., 2000). All 153 are used for evaluation. When the native sequences are folded with RhoFold+, the median GDT-TS is 0.6400 for the training set and 0.6014 for the evaluation set.

Train–test separation. The two sets are built independently and share no PDB entry. Character-wise and MD5 checks report no exact match between the 153 normalized evaluation sequences and the 527 training sequences. To characterize the residual similarity that remains after exact-duplicate removal, we compute each evaluation target's maximum global identity to the training set under Needleman–Wunsch alignment (Needleman & Wunsch, 1970). The median is 0.6207 and the mean 0.6615; 26 of the 153 targets reach 0.80, including 12 that reach 0.95; at the pair level, however, only 64 of the 80,631 evaluation–training pairs (0.08%) reach 0.80. All 153 targets are kept, with no exclusion based on similarity or results. Both policies in Table 4 are trained on the same training set and evaluated on the same evaluation set, so the reward comparison holds this residual similarity fixed.

Reference use. Backbones and masks condition generation. Native RNAfold pairings and nucleotide frequencies provide training references. Pairing evaluation uses base pairs extracted from the native 3D structures, available for 137 targets; the other 16 short RNAs have no pairing reference and remain in the tertiary, recovery, and thermodynamic evaluations.

D Implementation and evaluation protocol

Shared settings are 50 sampling steps (each predicts x_0 and re-noises it) at sampling temperature 1, bounded to $[0.2, 1.5]$ and raised by 0.1 for the single resampling round, AdamW (Loshchilov & Hutter, 2019) with zero weight decay, gradient clipping at 1, $K_t = 8$, $\beta = 0.1$, and $\lambda_{\text{ref}} = 0.02$. All reported evaluations use the same RhoFold+ weights.

Table 6: **Settings of the reported run.**

Setting	Value
Base / parameters	Pretrained RIDE / 10,214,213
Training / evaluation targets	527 / 153
Epochs / updates	3 / 390
Targets / candidates per update	4 targets $\times$ 8 candidates = 32
Learning rate / minimum ratio	2×10^{-4} / 0.3; 5% linear warmup, then cosine
Hardware / synchronization	Four GPUs / distributed gradient averaging
Held-out evaluation	Every 10 updates, from 0 to 390; 40 observations
Evaluated checkpoints / reported	180, 230, 340, 390 / 390 (last checkpoint)
Checkpoint selection	None: no validation split; the last checkpoint is reported

D.1 Aggregation and valid observations

For the main structural and pairing comparisons, candidate index 0 is taken for each target before aggregation. The index is fixed before scoring and is the same for every model, so each mean averages exactly one sample per target, with no selection. If $\mathcal{T}_m$ is the valid-target set for metric m, the reported mean is

$$\widehat{\mu}_m = \frac{1}{|\mathcal{T}_m|} \sum_{j \in \mathcal{T}_m} m_{j,0}. \tag{12}$$

The evaluation set has 153 targets, with 126 valid C1′ TM measurements. No best-of-eight selection or within-target eight-candidate average is used in Tables 1 and 2. The candidate-0 protocol yields one observation per target, so these aggregates are point estimates. Checkpoint means and standard errors (Appendix E) come from an earlier eight-candidate evaluation and are reported separately (Agarwal et al., 2021). The thermodynamic evaluation covers all 153 targets under its own protocol (Appendix D.3).

D.1.1 Secondary-structure valid counts

Table 7: **Valid target counts for the candidate-0 pairing metrics.** Counts correspond directly to Table 2; undefined measurements are excluded rather than replaced with zero. MORAD is update 390.

Model	n(Precision)	n(Recall)	n(F1)	n(MCC)
RiboDiffusion	133	137	137	133
gRNAde ($T = 0.1$)	132	137	137	132
RDesign	135	137	137	135
RIDE	133	137	137	133
MORAD	136	137	137	136

Recall and F1 have 137 valid targets for every model. Precision and MCC use 136 for MORAD, 133 for RIDE, 132 for gRNAde ($T = 0.1$), 135 for RDesign and 133 for RiboDiffusion. Precision and MCC are undefined for a design with no predicted base pair, whereas Recall and F1 remain defined; Recall and F1 therefore compare all models on the same 137 targets, and MORAD leads both (Table 2). For Precision and MCC, MORAD has the most valid targets.

D.1.2 Sequence diversity across eight candidates

Hamming diversity summarizes within-target normalized sequence differences. MORAD has 0.2344/0.3673, intermediate between gRNAde ($T = 0.1$)'s 0.1579/0.2406 and RIDE's 0.2971/0.5270. RDesign has the highest values, 0.4368/0.6184. These eight-candidate metrics are computed separately from the candidate-0 structural comparison and from the diversity curve recorded during training (Figure 5).

Table 8: **Sequence-diversity comparison.** Each model uses 153 targets and 1,224 candidates (eight per target). Bold marks the highest reported diversity; higher diversity alone does not establish greater structural accuracy.

Model	Hamming diversity ↑	3-mer diversity ↑	Targets	Candidates
RiboDiffusion	0.2744	0.3964	153	1224
gRNAde ($T = 0.1$)	0.1579	0.2406	153	1224
RDesign	**0.4368**	**0.6184**	153	1224
RIDE	0.2971	0.5270	153	1224
MORAD	0.2344	0.3673	153	1224

D.2 Metric definitions and interpretation

Structure and pairing. C4′ RMSD is $\sqrt{L^{-1}\sum_i d_i^2}$ for aligned residue distances d_i. GDT-TS uses the fractions below 1, 2, 4, and 8 Å (Zemla, 2003); our implementation averages them under one index-preserving Kabsch alignment (Kabsch, 1976). The C4′ TM-score averages $[1 + (d_i/d_0)^2]^{-1}$ after index-preserving Kabsch alignment, with the standard length-dependent d_0 (Zhang & Skolnick, 2004). C1′ TM uses US-align (Zhang et al., 2022). Precision is $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$, recall is $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, and F1 is $2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$; MCC applies the correlation coefficient to the base-pair confusion matrix (Matthews, 1975). EternaFold predictions cover canonical pairings (Wayment-Steele et al., 2022); the structure-derived reference base pairs include an estimated 12.8% noncanonical pairs. The native sequences score F1/MCC of 0.6652/0.6713 under this evaluation.

Additional structural diagnostics. Table 1 also reports success rate, sequence recovery, C4′ lDDT, pLDDT and a coarse clash score. lDDT measures local distance agreement, whereas pLDDT is the predictor's confidence, reported on a 0–1 scale (e.g., 0.6726 for MORAD). The coarse clash score is a coarse-grained steric-clash measure from the evaluation pipeline, not an all-atom clashscore. The evaluation report does not state the success threshold, lDDT distance cutoffs, or clash cutoff, so these values are reported as given. RiboDiffusion leads pLDDT and recovery despite lower alignment scores, showing that these metrics measure different properties.

Omitted outputs. We omit three further outputs: Rfam covariance-model scores, because the model library was unavailable; target probability, because the implementation returned values above one; and INF-wc, because it duplicates MCC. No zeros or values from other models are substituted for them.

Target compatibility and ensemble concentration. Let $P_i(s)$ denote nucleotide i's pairing-state probability, including the unpaired state, and $s_i^\star$ its target state. Ensemble defect and its normalized form are

$$\mathrm{ED} = \sum_{i=1}^{L} [1 - P_i(s_i^\star)], \qquad \mathrm{NED} = \mathrm{ED}/L. \tag{13}$$

NED is dimensionless and supports comparisons across RNA lengths; ED counts expected incorrect nucleotides (Zadeh et al., 2011). PE summarizes uncertainty in per-position pairing states; it is reported as computed by the evaluation pipeline, whose logarithm base was not recorded.

MFE is the energy of the minimum-energy secondary structure; EFE is the free energy of the Boltzmann ensemble. For one sequence and a consistent energy model, $\mathrm{MFEfreq} = \exp[-(\mathrm{MFE} - \mathrm{EFE})/(RT)]$ (McCaskill, 1990; Lorenz et al., 2011). The table reports each measured mean directly rather than deriving a mean frequency from mean energies. Lower PE and higher MFEfreq indicate a more concentrated ensemble; NED directly assesses target compatibility. Lower modeled energy alone does not establish target accuracy or folding kinetics.

D.3 Complete thermodynamic evaluation

EDiv measures the mean distance between secondary structures within the Boltzmann ensemble, not diversity among generated sequences (Lorenz et al., 2011). Larger values mean the ensemble contains more varied structures, which need not improve target accuracy, so EDiv is shown for context and

Table 9: **Complete thermodynamic measurements on 153 targets.** MFE/EFE: kcal/mol; ED: nt; NED: dimensionless; EDiv: mean base-pair distance; PE: positional entropy; MFEfreq: probability; T_m: °C; GC: fraction. MORAD is update 390. Bold marks the best of the six models in the indicated direction; Native is a reference and is not ranked. Dashes denote unreported reference values; $\sim$ marks metrics without a preferred direction, shown for context.

Metric	Ribo Diffusion	gRNAde $T = 0.1$	gRNAde $T = 0.8$	RDesign	RIDE	MORAD	Native
MFE $\sim$	-21.716	-28.092	-25.186	-28.863	-23.614	-30.149	-21.261
EFE $\sim$	-22.847	-28.918	-26.212	-29.708	-24.643	-30.937	—
ED $\downarrow$	43.653	26.913	28.486	25.502	29.184	**21.403**	28.049
NED $\downarrow$	0.6056	0.3458	0.3830	0.3349	0.3906	**0.2830**	0.3741
EDiv $\sim$	10.427	7.471	8.252	7.187	7.512	5.605	8.685
PE $\downarrow$	0.4284	0.2652	0.3260	0.2775	0.3059	**0.2102**	0.3581
MFEfreq $\uparrow$	0.2980	0.4289	0.3266	0.3961	0.3350	**0.4402**	0.3119
T_m $\sim$	73.91	73.07	73.57	71.71	72.54	72.52	—
GC $\sim$	0.5734	0.5529	0.5650	0.5195	0.5400	0.5393	—

not ranked. MORAD has EDiv 5.605 versus RIDE's 7.512, consistent with a more concentrated ensemble. T_m (predicted melting temperature) and GC content are reported as descriptors: MORAD has 72.52 °C and GC 0.5393, versus 72.54 °C and 0.5400 for RIDE. The native reference has ED 28.049 and NED 0.3741; it is a sequence reference rather than a guaranteed optimum under the evaluation model.

E CHECKPOINT COMPARISON AND ADDITIONAL RESULTS

Update 390 is the last checkpoint and leads all five means in this evaluation. Its eight-candidate GDT/TM values (0.3774/0.3338) are kept here for reference; the main comparison uses the candidate-0 values (0.3827/0.3400).

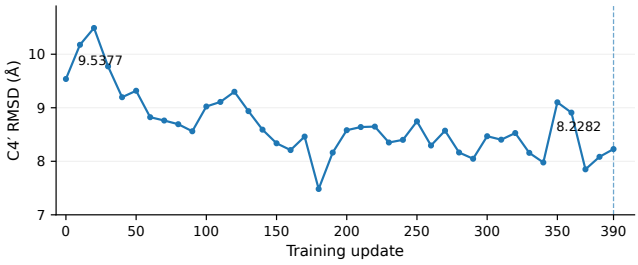

Figure 4: **Complete RMSD trajectory.** All 40 held-out measurements are connected in time order. Lower RMSD indicates closer agreement after alignment; the vertical marker shows update 390.

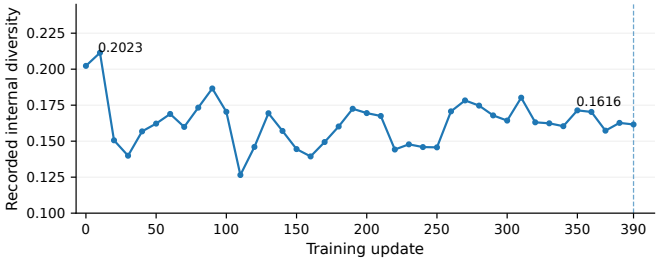

Figure 5: **Within-target diversity during training.** Diversity decreases from 0.2023 to 0.1616. The eight-candidate comparison across models is in Table 8.

Table 10: **Checkpoint comparison.** Means over eight candidates per target, with standard errors, for four checkpoints from an earlier evaluation run; they differ from the candidate-0 results in Tables 1 and 2, which also report update 390, the last checkpoint.

Checkpoint	GDT-TS ↑	TM(C4′) ↑	RMSD (Å) ↓	C1′ TM ↑	Recovery ↑
Step 180	0.3682 ±0.0177	0.3218 ±0.0198	8.263 ±0.490	0.3716 ±0.0171	0.5074 ±0.0108
Step 230	0.3617 ±0.0184	0.3208 ±0.0201	8.733 ±0.554	0.3697 ±0.0176	0.5021 ±0.0107
Step 340	0.3686 ±0.0185	0.3271 ±0.0205	8.435 ±0.530	0.3771 ±0.0177	0.5032 ±0.0104
MORAD (390)	0.3774 ±0.0179	0.3338 ±0.0202	7.981 ±0.489	0.3833 ±0.0177	0.5102 ±0.0106

E.1 Training diagnostics

Figures 4 and 5 show all 40 evaluations run during training. Over these evaluations, recovery changes from 0.5020 to 0.5069 and success rate from 0.2288 to 0.3137.

Smoothing. For a raw structural score q_j at update $10j$, $j = 0, \ldots, 39$, Figure 3 displays

$$\widetilde{q}_j = \frac{1}{|W_j|} \sum_{k \in W_j} q_k, \qquad W_j = \{\max(0, j-2), \ldots, \min(39, j+2)\}. \tag{14}$$

All three structural scores use this window. The raw values remain visible in the figure; the diagnostics in this appendix are not smoothed.