Pareto Preference Optimization for Structure- and Stability-Aware RNA Inverse Folding

Lawrence Minghao Sun[1*] **Hanqun Cao**[2*] **Zhou Zhang**[3*] **Chen Wei**[4] **Liang Wang**[1,6]
Tianrui Jia[1] **Zhiyuan Liu**[1] **Tianfan Fu**[3] **Xiangru Tang**[5] **Yejin Choi**[7]
Pheng-Ann Heng[2] **Fang Wu**[7†] **Yang Zhang**[1†]

[1]National University of Singapore [2]The Chinese University of Hong Kong
[3]Nanjing University [4]Xi'an University of Posts & Telecommunications [5]Yale University
[6]Institute of Automation, Chinese Academy of Sciences [7]Stanford University

Abstract

RNA inverse folding seeks sequences that fold into a desired three-dimensional structure. Learning to recover native sequences alone does not directly optimize structural fidelity and thermodynamic stability, while optimizing individual physical proxies can favor compositional shortcuts. We present RIBOPO, a preference-optimization framework for structure- and stability-aware RNA inverse folding. RIBOPO constructs preference pairs through ε-Pareto dominance with variability-scaled margins, selecting sequences that jointly improve structural and thermodynamic scores. A structural quality gate constrains the preferred candidates, and frozen-reference optimization with a decreasing-margin curriculum refines the policy across rounds. A clipped importance-reweighting extension supports continued learning from offline candidate pools. On the DAS benchmark, a single default checkpoint improves secondary-structure consistency, three-dimensional designability, and thermodynamic scores over the pretrained inverse-folding model. Evaluation with independent structure predictors supports the structural gains, while multi-seed experiments establish the reproducibility of the quality-control and reference designs. Matched-utility reranking shows that the refined candidate pools reach the base model's joint structural success with fewer samples. Specialist variants provide additional operating points for applications that prioritize individual objectives.

https://github.com/minghao-sun-sc/ribopo

1 Introduction

The RNA inverse folding problem asks: which nucleotide sequence reliably folds to a specified 3D backbone? It is the cornerstone of rational RNA engineering for gene regulation [1], therapeutics [2], and synthetic biology [3]. Geometric neural networks have made significant progress [4–7], but a key gap remains: optimizing for sequence recovery or a single structural proxy ignores RNA's pronounced *ensemble* flexibility [8], leading to designs that score well on one axis but fail to consistently adopt the intended conformation under physiological conditions.

Designability, the density of sequences that fold into a target structure, is a critical functional property [9, 10]. Highly designable structures are mutation-robust, easier to optimize, and inhabit thermodynamically favorable basins [11]. Optimizing geometry alone may reach thin, fragile ridges in sequence space; coupling geometry with ensemble stability biases search toward robust basins. Yet

*Equal contribution.
†Corresponding authors: fangwu97@stanford.edu, zhang@nus.edu.sg.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

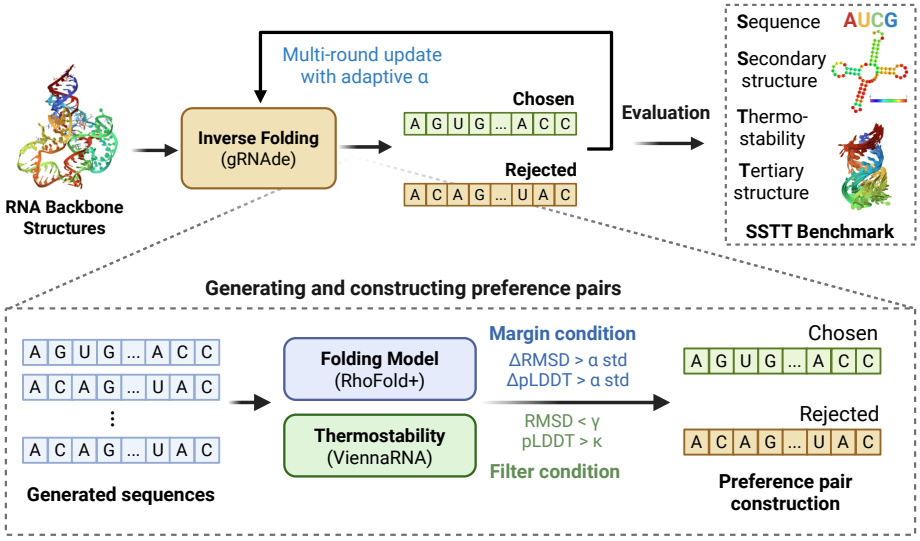


Figure 1: Overview of RIBOPO. (a) Candidate generation from π_{ref} with multi-proxy scoring (pLDDT, RMSD, MFE). (b) ε-Pareto-dominance preference set with structural quality gate. (c) Frozen-reference DPO with decreasing-margin curriculum.

existing benchmarks emphasize global geometric correspondence (TM-score [12], RMSD, pLDDT [13]) and provide limited insight into ensemble viability [14, 15].

We present RIBOPO, a preference-optimization framework that addresses two RNA-specific obstacles: heterogeneous scales across physical proxies and sequence–structure degeneracy that enables compositional shortcuts through GC enrichment. The framework selects jointly improving sequence pairs, applies a structural quality gate, and fine-tunes a pretrained inverse-folder against a fixed reference. This connects multi-objective preference construction to stable sequence refinement without choosing a scalar reward that can compensate for structural deterioration with lower free energy.

Contributions.

- **Variability-aware ε-Pareto preference construction**, with labels consistent with every nonzero nonnegative scalarization and a structural quality gate for preferred sequences (§3.2).
- **Frozen-reference multi-round preference optimization**, supported by a fixed-reference variational characterization and a pair-distribution update bound; clipped reweighting improves extended-round training (§3.3).
- **Reproducible control of compositional shortcuts**, examined through component ablations, multi-seed training, and GC/length-adjusted thermodynamic analysis (§4.4–§4.5).
- **SSTT evaluation across sequence, secondary structure, tertiary structure, and thermostability**, with a single default checkpoint and independent structure predictors (§4.1–§4.2).

On the 98-target DAS benchmark, the default RIBOPO-R2 checkpoint increases EternaFold scMCC from 0.606 to 0.690 and the fraction of designs with RMSD at most 8 Å from 41.6% to 51.0%. The same checkpoint improves RMSD and MFE, demonstrating coordinated gains across structural and thermodynamic evaluation axes. Independent-predictor evaluation and multi-seed ablations provide complementary evidence for these improvements. We evaluate specialist operating points and a weight-conditioned extension to characterize additional design trade-offs, while keeping R2 fixed for the primary comparison.

2 Related Work

RNA structure prediction and inverse folding. Tertiary RNA predictors span energy-function methods [16, 17], EvoFormer architectures [18–20], and complex models [21, 22]; secondary scorers include ViennaRNA [23] and EternaFold [24]. Existing inverse-folders use GNN encoders [25,

26, 7, 27–29, 5], discrete-diffusion bases [30, 6], or MPNN adaptations [31, 32]: gRNAde extends single-state to multi-state design and RhoDesign reports experimental validation. None directly optimize a joint structure–ensemble objective: sequences are recovered by maximum-likelihood imitation of natives and inherit its biases.

Preference optimization and RL for biomolecular design. RL and preference optimization on top of pre-trained inverse-folders have been explored for protein design: ProteinZero [33], ResiDPO [34], PLM-RL [35], peptide DPO [36], RL on discrete-diffusion bases (RL-DIF [37], DRAKES [38]), reward-guided search (ProtInvTree [39]), gradient-free finetuning (GLID^{2}E [40]), and energy-based Markov-bridge variants (EnerBridge-DPO [41]). The closest work, ProtAlign [42] on ProteinMPNN, uses MODPO scalarization [43]; RIBOPO instead uses scalarization-agnostic ε-Pareto labels with a fixed reference and pair-distribution analysis (Propositions 1–2). To our knowledge, no prior work demonstrates effective preference-based fine-tuning on RNA inverse folding; RIBOPO confronts three RNA-specific obstacles distinct from protein DPO: noisy 3D oracles whose error correlates with composition, sequence–structure degeneracy admitting compositional reward hacking, and a candidate-pool ceiling that pushes multi-round training into a regime where off-policy bias is non-negligible.

3 Methods

3.1 Problem and Multi-Round Setup

We frame RNA inverse folding as a multi-objective design problem over $\mathcal{S} = \{A, U, C, G\}^L$ with a vector-valued quality function $\phi(s, \mathcal{G}) \in \mathbb{R}^M$ aggregating 3D fidelity (pLDDT, RMSD) and ensemble stability (MFE); the goal is to find sequences that Pareto-dominate the candidate pool. The autoregressive policy $\pi_\theta(s \mid \mathcal{G}) = \prod_i \pi_\theta(s_i \mid s_{<i}, \mathcal{G})$ uses the GVP-GNN backbone of gRNAde [5] and is trained against a frozen reference π_{ref} initialized from the gRNAde checkpoint. Optimization proceeds in R rounds with offline preference sets, a frozen reference, and a decreasing-margin curriculum (§3.2–§3.3); the resulting ε-Pareto pair construction is consistent with every nonzero nonnegative scalarization of ϕ at the label level (§3.2).

3.2 Preference Construction under Heterogeneous Noisy Proxies

RNA preference learning differs from LLM preference learning in two respects. First, candidates are scored by *multiple* physical proxies (3D fidelity, secondary-structure consistency, free energy) with very different noise scales rather than a single learned reward model. Second, RNA's sequence-structure degeneracy [11] creates exponentially large neutral networks in which compositional optimization (e.g., GC enrichment for MFE) trivially exploits any single proxy. These two properties motivate the construction below.

ε-Pareto-dominance preference set. Let $\phi(s, \mathcal{G}) = (\phi_1, \ldots, \phi_M) \in \mathbb{R}^M$ be the per-metric quality vector, oriented so that larger is preferred ($M = 3$ with $\phi = (\text{pLDDT}, -\text{RMSD}, -\text{MFE})$). EternaFold scMCC, $P(S_0)$, and ensemble defect are not used in preference labels or the structural gate; they are label-excluded evaluation metrics. We additionally use structure predictors absent from preference construction to assess transfer across evaluators. For per-coordinate margins $\varepsilon_r \succeq 0$ (instantiated as $\varepsilon_{r,m} := \alpha_r \sigma_m$ below), define

$$\phi(s^w) \succeq_{\varepsilon_r} \phi(s^l) \iff \forall m : \phi_m(s^w) - \phi_m(s^l) > \varepsilon_{r,m}. \tag{1}$$

At round r, $\mathcal{D}_r = \{(s^w, s^l, \mathcal{G}) : \phi(s^w, \mathcal{G}) \succeq_{\varepsilon_r} \phi(s^l, \mathcal{G})\}$, augmented with a structural quality gate on the winner: $\text{pLDDT}(g(s^w)) > \kappa$ and $\text{RMSD}(\mathcal{G}, g(s^w)) < \gamma$, with $(\kappa, \gamma) = (0.70, 8.0\,\text{Å})$. The gate requires preferred sequences to satisfy both structural thresholds, complementing the relative improvements enforced by Pareto dominance (§4.4).

Why Pareto, not scalarization (label-level guarantee). A preference $\phi^a \succeq_\varepsilon \phi^b$ satisfies $w^\top \phi^a > w^\top \phi^b + w^\top \varepsilon$ for every $w \succeq 0$ with $w \neq 0$, so each retained label is consistent with every nonzero nonnegative scalarization simultaneously. Conversely, the canonical GC-hacking margin $(\Delta\text{pLDDT}, \Delta\text{RMSD}, \Delta\text{MFE}) = (-0.05, +1\,\text{Å}, -8)$ is accepted by any scalarization with sufficient weight on $-$MFE but rejected by ε-Pareto for any $\varepsilon \succ 0$. This label-level property does

not certify that the learned policy is globally Pareto-monotonic on unseen sequences. The construction differs from soft scalarization (MODPO [43]) in that pair labels are scalarization-agnostic by construction rather than implicitly determined by a chosen weight vector.

Variability-aware margin. The three RNA proxies have different empirical scales (candidate-pool SDs of approximately 0.04 for pLDDT, 1.5 Å for RMSD, and 4 kcal/mol for MFE). We set $\varepsilon_{r,m} = \alpha_r \sigma_m$ with $\alpha_{1,2} = 0.25$ and $\alpha_{3,4,5} = 0.125$. Scaling by pool variability makes pair selection invariant to positive changes of metric units. Decreasing α_r implements a refinement curriculum: strongly separated pairs early, followed by finer improvements in later rounds. Here σ_m measures candidate variability; Appendix A gives the corresponding selection properties.

Thermodynamic-surplus filter. For the specialist thermodynamic-surplus variant, we fit candidate-pool MFE on length and GC content, $\text{MFE} = a + bL + c(\text{GC} \cdot L)$, and additionally require the winner to improve the fitted MFE residual. This filter removes 23.7% of pairs at $\alpha = 0.25$ and selects thermodynamic improvements that persist after this adjustment. It is an optional operating point, separate from the default R2 configuration.

3.3 Pareto-Constrained DPO with Anchored-Reference Curriculum

Given $\mathcal{D}_r$, we fit an implicit reward via the DPO loss [44]:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(s^w, s^l, \mathcal{G}) \sim \mathcal{D}_r}\Big[\log \sigma\Big(\beta \log \frac{\pi_\theta(s^w \mid \mathcal{G})}{\pi_{\text{ref}}(s^w \mid \mathcal{G})} - \beta \log \frac{\pi_\theta(s^l \mid \mathcal{G})}{\pi_{\text{ref}}(s^l \mid \mathcal{G})}\Big)\Big]. \tag{2}$$

We add an SFT anchor on the chosen sequences with weight λ_{SFT}=0.10 to prevent the known low-margin failure mode in which DPO depresses $\pi_\theta(s^l)$ rather than concentrating mass on s^w [45]. We use $\beta = 0.12$ for the primary configuration. We sum sequence log-probabilities in the pairwise logit; candidates within each backbone have equal length.

Frozen reference and variational characterization. We keep π_{ref} fixed at the gRNAde initialization across rounds, retaining a common distributional anchor for DPO. The standard KL-regularized formulation underlying DPO [44] makes this role explicit.

Proposition 1 (Fixed-reference variational characterization) *For a fixed backbone, let q be a reference distribution with full support on the finite sequence space, let u be a fixed bounded reward, and let $\beta > 0$. Define $J_q(\pi) = \mathbb{E}_\pi[u] - \beta \text{KL}(\pi \| q)$ and $\pi^*(s) = q(s) e^{u(s)/\beta}/Z$, where Z normalizes the distribution. Then*

$$J_q(\pi^*) - J_q(\pi) = \beta \text{KL}(\pi \| \pi^*), \qquad \text{KL}(\pi^* \| q) \le \frac{\sup u - \inf u}{\beta}.$$

The proof is in App. B. We assess the stability of frozen-reference training through matched multi-seed ablations (§4.4).

Distribution shift under static pairs. Preference pairs are drawn from the reference candidate pool, while the policy changes during training. The following result characterizes the resulting difference in expected per-pair updates.

Proposition 2 (Pair-distribution update discrepancy) *At fixed θ, let $g_\theta(z)$ be the per-pair loss gradient and $\Gamma(\theta; \nu) = \mathbb{E}_{z \sim \nu}[g_\theta(z)]$. Let $P = \nu_{\text{ref}}$ and $Q = \nu_\theta$ be pair distributions under the same preference rule. If $Q \ll P$, $\|g_\theta(z)\|_2 \le G$, and $D_2(Q\|P) = \log \mathbb{E}_P[(dQ/dP)^2] < \infty$, then*

$$\|\Gamma(\theta; Q) - \Gamma(\theta; P)\|_2 \le G \sqrt{e^{D_2(Q\|P)} - 1}.$$

This change-of-measure bound motivates reweighting during pair reuse. We use a geometric-mean, clipped, self-normalized weighting rule (Eq. 3); its extended-round performance is evaluated in §4.6. The proof and estimator are given in App. C.

3.4 Stage 2: Weight-Conditioned Policy via FiLM (Optional)

A scalarization $f(\phi) = w^\top \phi$ with $w \in \Delta^{M-1}$ specifies relative preferences over the objectives. To allow inference-time control over the operating point without retraining, we add a FiLM head [46] to the gRNAde decoder: scalar features h are modulated by $h \to h(1 + \gamma(w)) + \beta(w)$, with $\gamma, \beta : \mathbb{R}^M \to \mathbb{R}^{h_{\text{dim}}}$ initialized to zero (residual initialization makes the wrapped policy exactly equivalent to the base at w unspecified). Total cost: 1,024 extra parameters. During training we sample $w \sim \text{Dirichlet}(1)$ per batch; at inference, querying at $w = (1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$, or the centroid exposes a parameterized set of operating points. We empirically find this set to be *discrete* rather than continuous (§4.8): vertex queries converge to a single thermo-strong mode while the centroid produces a distinct 2D-strong mode. Details are in Appendix D.

3.5 SSTT Evaluation Panel

We evaluate on **SSTT**: **S**equence (recovery and 3-mer profile deviation), **S**econdary structure (EternaFold scMCC), **T**ertiary structure (TM, GDT, RMSD, pLDDT, lDDT, INF [47], clash [48]), and **T**hermostability (MFE, ensemble defect, target-structure probability $P(S_0)$, Shannon entropy, melting temperature). The reported 3-mer statistic compares each design's normalized 64-dimensional frequency profile with that of the native sequence; it measures local sequence-profile deviation, rather than diversity among generated candidates. We distinguish optimization-aligned metrics (RhoFold+ pLDDT/RMSD and ViennaRNA MFE), label-excluded metrics such as scMCC, and evaluation with predictors absent from training.

4 Experiments

Setup. We use the released gRNAde DAS split unchanged: 4,025 training, 100 validation, and 98 test structures [49, 5]. It assigns entire structural clusters to splits using a US-align/qTMclust cutoff of 0.45; stored train and test cluster identifiers are disjoint. Preference pools contain eight sequences per backbone, with 20,811/505/429 train/validation/test pairs after length-consistency filtering. Evaluation uses EternaFold for secondary structure, RhoFold+ for tertiary structure, and ViennaRNA for thermodynamics. Unless stated otherwise, results average eight designs per target at temperature 0.1. Split details and optimization settings are in Appendix F.

Default checkpoint and operating points. **RIBOPO-R2 is the default checkpoint**, obtained after two rounds from the canonical upstream gRNAde model. We use R2 for the primary DAS and independent-predictor comparisons. R4 and thermodynamic-surplus are specialist operating points for applications that specify a secondary-structure or thermodynamic objective in advance. The Stage-1/Stage-2 variants and the five-round IS experiments are reported separately; their results are not combined into the default checkpoint's headline performance.

Baselines. We compare with gRNAde and six vanilla inverse-folders: RhoDesign [7], RIdiffusion [6], RiboDiffusion [30], NA-MPNN [31], RDesign [25], and R3Design [26]. Table 1 isolates preference optimization on the shared gRNAde backbone and split. External-baseline results on their respective overlap-audited subsets are reported in Table 2, with full-set results in Appendix E. These baseline-specific subsets provide descriptive context; the direct R2–gRNAde comparison uses the same 98 targets.

4.1 Main Results

RIBOPO-R2 improves structural consistency and thermodynamic scores within one checkpoint (Table 1). EternaFold scMCC increases from 0.606 to 0.690, while RMSD decreases from 11.45 to 10.23 Å. The fraction of designs meeting the 8 Å RMSD threshold rises by 9.4 percentage points, from 41.6% to 51.0%. MFE shifts from -28.64 to -31.91 kcal/mol. Recovery changes from 0.529 to 0.502 as optimization prioritizes structure and stability over reproducing the native sequence. R4 provides a secondary-structure-focused operating point with scMCC 0.71; the default R2 retains the stronger balance of tertiary structure and thermodynamic scores.

The SSTT panel separates secondary-structure consistency from the metrics used in preference construction. In particular, the scMCC gain is obtained without using EternaFold scores in preference

Table 1: SSTT evaluation on the 98-target DAS test set. RIBOPO-R2 is the default checkpoint (shaded); other rows show training rounds or specialist operating points. All rows use the same gRNAde backbone and test targets. Best values are bold within each configuration block, except the descriptive 3-mer deviation statistic. Threshold columns report success fractions; RMSD is in Å and MFE in kcal/mol. 3D scores use RhoFold+; evaluation details are in Appendix F.

Model	Sequence		2D	3D					Thermo
	Rec. ↑	3-mer dev.	scMCC ↑	pLDDT ↑	pLDDT ≥0.7	RMSD ↓	RMSD ≤8Å	TM ↑	MFE ↓
Shared gRNAde backbone; default and round-wise operating points.									
gRNAde (base)	**0.529**	0.82	0.606	0.621	0.381	11.45	0.416	0.293	-28.64
RIBOPO (Round 1)	0.502	0.90	0.690	**0.660**	**0.470**	10.26	0.450	0.300	-31.03
RIBOPO-R2 (default)	0.502	0.87	0.690	0.650	0.450	**10.23**	**0.510**	**0.310**	**-31.91**
RIBOPO (Round 4)	0.506	0.89	**0.710**	0.650	0.420	10.41	0.440	0.290	-30.52
Specialist variants; additional loss and weight sweeps in Appendix J.									
RIBOPO + thermo-surplus	**0.496**	0.88	0.643	0.630	0.438	11.21	0.430	0.269	**-30.74**
RIBOPO (Stage-2, centroid)	0.490	0.85	**0.656**	0.601	0.360	12.30	0.417	0.240	-28.68
RIBOPO (Stage-1)	0.492	0.90	0.634	**0.645**	**0.450**	**10.41**	**0.480**	**0.285**	-30.65

Table 2: Baseline-specific overlap-audited results. Each external method uses its own cleaned DAS subset, so scores across those rows are descriptive rather than a common-target ranking. The lower block compares gRNAde and the default RIBOPO-R2 on the same 98 targets. Audit criteria are summarized in Appendix E.

Model	Rec.↑	scMCC↑	RMSD (Å)↓	RMSD≤8Å↑	TM↑	MFE (kcal/mol)↓
RDesign	0.42	0.21	16.48	0.23	0.13	-22.06
R3Design	0.51	0.59	14.52	0.47	0.27	-26.57
RhoDesign	0.47	0.21	14.82	0.18	0.14	-13.53
RiboDiffusion	0.52	0.49	13.52	0.25	0.22	-20.28
Shared DAS-98 targets						
gRNAde (base)	0.53	0.61	11.45	0.42	0.29	-28.64
RIBOPO-R2	0.50	0.69	10.23	0.51	0.31	-31.91

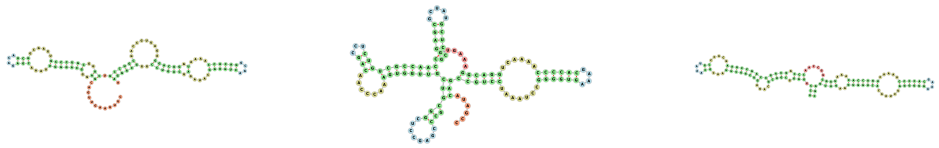
Table 3: Independent evaluation of fixed designs. Entries show gRNAde → RIBOPO-R2. N is the number of complete target pairs; all predictors use single-sequence inputs.

Predictor	N	RMSD (Å)↓	TM↑	pLDDT↑
Boltz-2	98	3.256 → 3.178	0.398 → 0.411	0.624 → 0.656
AlphaFold3	98	3.286 → 3.157	0.372 → 0.375	0.576 → 0.587
Protenix	93	3.186 → 3.029	0.399 → 0.414	0.638 → 0.649

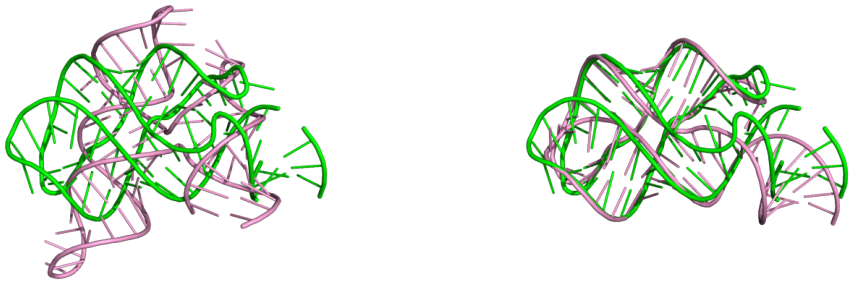
labels or the structural gate. The thermodynamic-surplus and Stage-2 rows characterize additional operating points rather than components of the default result. Target-ensemble probability and defect are reported as secondary, checkpoint-specific observations in Appendix I.

4.2 Evaluation with Independent Structure Predictors

We re-fold fixed, unre-ranked R2 and gRNAde sequences using Boltz-2 [50], AlphaFold3 [22], and Protenix [51], none of which is used to construct preferences. Table 3 shows structural improvements across complementary evaluators: Boltz-2 increases TM-score and confidence, while Protenix improves RMSD and TM-score. These changes are supported by paired tests at the 0.05 level; AlphaFold3 changes are in the same direction, with non-significant paired differences. These panels cover 98 targets for Boltz-2 and AlphaFold3 and 93 for Protenix. Predictor settings and paired tests are given in Appendix O.



(a) 4MEG_1_B: native-sequence prediction (left), gRNAde (middle), and RIBOPO (right).



(b) 5C7W_1_C: gRNAde (left; TM 0.060, RMSD 18.46 Å) and RIBOPO (right; TM 0.462, RMSD 3.68 Å). Both designs have recovery 0.446.

Figure 2: **Structural consequences of sequence design.** (a) EternaFold predictions for native and designed sequences, rendered with forna [52]. (b) AlphaFold3 predictions superposed with PyMOL [53] (native green, design pink). Cases were selected post hoc to illustrate the secondary-topology contrast and the tertiary-structure difference at matched sequence recovery.

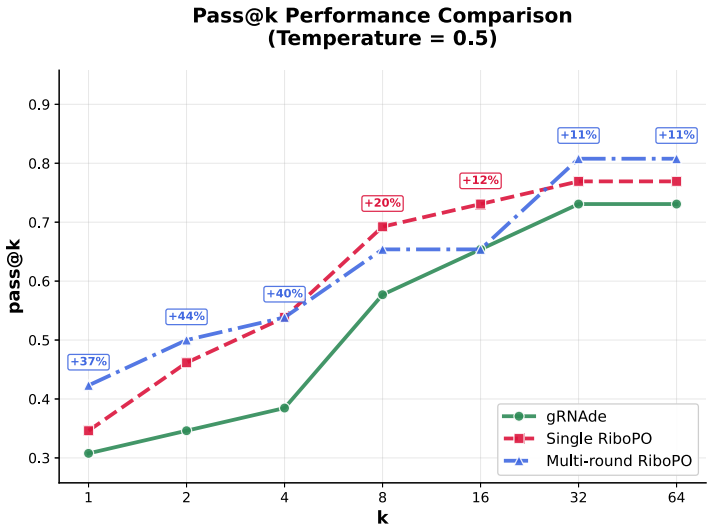
4.3 Structural Case Studies

The case studies illustrate why sequence agreement alone does not determine structural fidelity. For chain 4MEG_1_B, the gRNAde design's EternaFold prediction contains a four-way junction, whereas the RIBOPO design retains the extended topology of the native-sequence prediction (Figure 2a). For chain 5C7W_1_C, the two designs have identical recovery of 0.446, yet their AlphaFold3 predictions differ substantially: RMSD is 18.46 Å for gRNAde and 3.68 Å for RIBOPO (Figure 2b). The matched-recovery case shows that sequence agreement alone can overlook substantial differences in predicted three-dimensional structure.

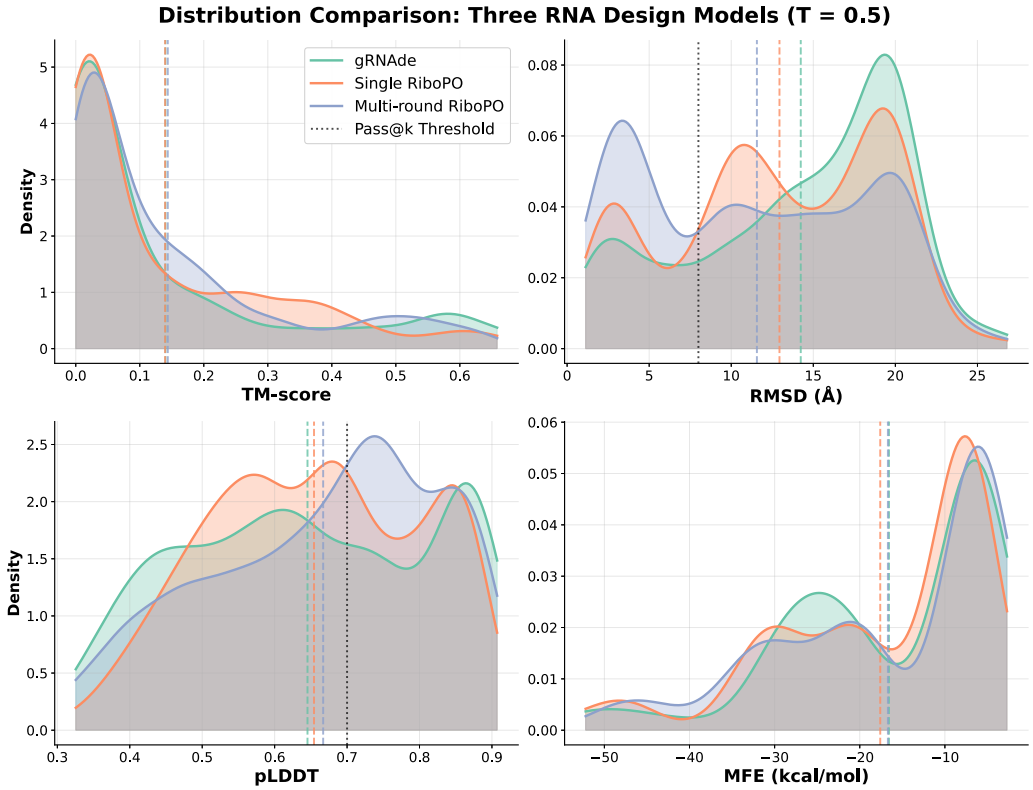
4.4 Reward-Hacking Ablation

Component ablations identify complementary roles for structural selection and reference anchoring. In the component-ablation runs, reference updating produces much lower MFE but worse RMSD, while the no-gate control favors lower free energy (Appendix G). Multi-seed replications reinforce the compositional signature: the MFE-only/no-gate control increases GC content at all three matched seeds, with MFE reductions of 4.3–7.9 kcal/mol relative to gated training. Recovery and scMCC alone do not reveal this shortcut, underscoring the need to evaluate structure and thermodynamics together.

Across five training seeds, frozen-reference recovery is 0.496 ± 0.006, compared with 0.484 ± 0.026 under reference updating; scMCC is also more consistent across seeds. The decreasing-margin curriculum gives a smaller, reproducible gain: mean scMCC is 0.666 versus 0.646 for constant margins in the two-seed matched evaluation. These replications use a shared reduced training schedule, summarized with aggregate and matched-seed results in Appendix H. Complementary DPO/SFT loss ablations are reported in Appendix R.



(a) Pass@k sampling efficiency.



(b) Metric distribution shifts.

Figure 3: **Sampling efficiency and distribution shifts.** Both panels use temperature 0.5. (a) Pass@k on 14 DAS structures under the joint pLDDT$\geq$0.70 and RMSD$\leq$8 Å criterion. (b) Metric distributions from 64 designs for each of 26 chains from 14 DAS complexes: multi-round RiboPO improves RMSD and MFE and reduces the low-confidence pLDDT tail, with greater mass above the feasibility threshold. The high-confidence pLDDT mode remains approximately unchanged. Black dotted lines mark the RMSD and pLDDT feasibility thresholds. The separate temperature-0.1 reranking comparison uses all 98 targets (§4.7).

4.5 GC-Controlled Regression and Sample Efficiency

Thermodynamic gains after GC/length adjustment. We assess composition sensitivity by regressing MFE on length and length-scaled GC content, MFE $= a + bL + c(\mathrm{GC} \cdot L)$. On the matched design panel used for this analysis, the reference-fitted model retains a residual contrast of -2.48 kcal/mol from the raw -3.57 kcal/mol contrast. Reference-only and pooled fits retain approximately 65–69% of that contrast, indicating that a substantial energy difference persists after the specified adjustment. The specialist thermodynamic-surplus filter selects pairs with improved residual MFE; its results are reported separately in Appendix P. Cross-engine evaluation with NUPACK and RNAstructure provides a further check on the thermodynamic score changes (Appendix N).

Per-sample success rate under a joint structural criterion. Figure 3a evaluates the probability of obtaining a design meeting the joint structural criterion as the sampling budget increases. At temperature 0.5, RIBOPO has higher success across the plotted budgets, with the largest relative gains in the small-budget regime. The distributional view in Figure 3b clarifies the pLDDT improvement: for multi-round R2, the fraction below 0.50 decreases from 24.0% to 17.9%, and the median rises from 0.637 to 0.699. The improvement is concentrated in the lower tail and around the feasibility threshold rather than an upward shift of the 0.85–0.90 mode. Appendix I summarizes target-level analyses; the full-DAS matched-utility comparison below examines selection from each model's candidate pool.

4.6 Multi-round preference learning under off-policy bias

Static-pair reuse introduces a mismatch between the reference-induced pair distribution and the current policy; Proposition 2 characterizes its effect on the expected update. We examine a practical reweighting extension that takes the geometric mean of sequence likelihood ratios, clips the result, and normalizes weights within each batch (Appendix C). In the five-round comparison, the static-pair run's RMSD success rate falls from 0.481 to 0.246, whereas the clipped-IS run finishes at 0.486 (Figure 4). Matched-seed comparisons show a consistent recovery benefit: static reuse reduces recovery at both training seeds, whereas clipped-IS preserves or slightly improves it. These results support reweighting for extended-round refinement, alongside the two-round default configuration. Full trajectories and weight diagnostics are in Appendices H and L.

4.7 Same-pool reranking control

We apply the same selection utility separately to each model's candidate pool to measure the candidate budget needed for a successful design. Under Pareto-feasible-then-max-$P(S_0)$ selection, both R2 checkpoints from the dedicated static-pair and clipped-IS experiments match gRNAde's best-of-64 joint success with eight candidates. The shared selector isolates candidate-pool quality from the choice of reranking rule. This eightfold candidate-budget comparison uses the extension runs, reported separately from the default R2 checkpoint in Table 1; Appendix M gives the matched comparison.

4.8 Beta Sweep, Loss Ablation, and Stage-2 Operating Points

A β-sweep and alternative-loss ablation (IPO [54], KTO [55], Pareto-DPO Stage 1, Stage 2) on a matched pair set and budget (Appendix J) shows method specialization on different Pareto axes: **Stage-2 centroid** reaches scMCC 0.656 on 2D among the alternative single-policy variants (multi-round R=4 at 0.71 is higher overall, Table 1); **Stage-1** (stochastic Dirichlet w, no FiLM) has the strongest 3D metrics among these alternatives (RMSD 10.41 Å, TM 0.285, RMSD$\leq$8Å success fraction 0.480); **thermo-surplus** provides a 2D–thermo operating point ($P(S_0)$ 0.0155, T_m 28.4°C). At β=0.001, GC enriches to 0.694, RMSD increases to 14.7 Å, and TM decreases to 0.194, illustrating the sensitivity of the structural–thermodynamic balance to the preference scale. IPO and KTO underperform DPO; KTO trades 3D fidelity for 2D consensus. The three operating points (Stage 1 / Stage 2 / thermo-surplus) illustrate distinct trade-offs on the (scMCC, RMSD, $P(S_0)$) axes.

Stage-2 operating points. The weight-conditioned extension yields two empirical modes: a centroid query emphasizes secondary-structure consistency, while vertex queries favor thermodynamic scores (Figure 5). Pareto-dominant labels preserve their ordering across nonnegative scalarizations, so they

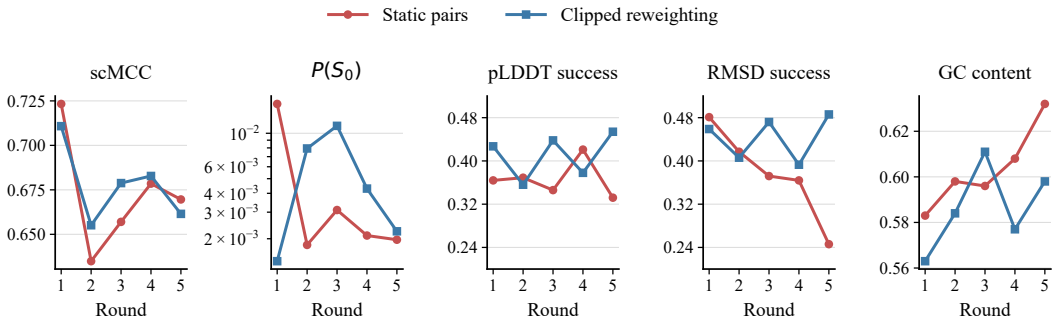


Figure 4: **Extended-round preference optimization.** Matched-seed trajectories for static-pair DPO (red) and clipped reweighting (blue). Reweighting retains stronger structural success at later rounds. Additional matched training seeds support improved recovery retention (Appendix H); likelihood and weight diagnostics are in Appendix L.

do not supply weight-dependent preference reversals. The observed modes characterize this optional extension; the default R2 model does not require inference-time weight selection.

5 Discussion

RIBOPO demonstrates that preference optimization can improve structure-conditioned RNA sequence design across complementary evaluation axes. Its central design choice is to learn from jointly improving pairs while maintaining structural feasibility and a fixed reference. The default R2 checkpoint provides a single balanced operating point; independent structure predictors and repeated training runs strengthen the evidence for the structural improvements and reference design. Specialist configurations and reweighting extend the framework to different design priorities and longer optimization horizons.

The theoretical results characterize scalarization-consistent labels, the fixed-reference objective, and pair-distribution mismatch. The experiments establish how these choices behave in the RNA-design setting. A useful next direction is refreshed candidate generation, which can supply new preference pairs as the policy evolves.

Scope and impact. The evaluation measures computational structure and thermodynamic proxies; prospective experimental validation remains necessary for application-specific folding and function. Sequence recovery reflects a distinct objective, and GC/length adjustment addresses the specified compositional effects. RIBOPO is intended to support rational RNA design for user-supplied backbones, including riboswitches, ribozymes, and synthetic regulatory elements.

Acknowledgments and Disclosure of Funding

This work was supported in part by the Ministry of Education, Singapore (MOE-T1251RES2309, MOE-T2EP20125-0014, and MOE-T2EP20225-0028), the Agency for Science, Technology and Research (A*STAR), Singapore (IAF-PP H25J6a0034), and the National Research Foundation, Singapore (NRF-CRP33-2025-0003 and AI4SCT_2025-0014). The authors declare no competing interests.

References

[1] Alexander A Green, Pamela A Silver, James J Collins, and Peng Yin. Toehold switches: de-novo-designed regulators of gene expression. *Cell*, 159(4):925–939, 2014.

[2] Theresa Yip, Xiaodong Qi, Hao Yan, and Yung Chang. Therapeutic applications of RNA nanostructures. *RSC Advances*, 14(39):28807–28821, 2024.

[3] Blaine A Pfeifer, Marie Beitelshees, Andrew Hill, Justin Bassett, and Charles H Jones. Harnessing synthetic biology for advancing RNA therapeutics and vaccine design. *npj Systems Biology and Applications*, 9(1):60, 2023.

[4] Fang Wu, Lirong Wu, Dragomir Radev, Jinbo Xu, and Stan Z Li. Integration of pre-trained protein language models into geometric deep learning networks. *Communications Biology*, 6 (1):876, 2023.

[5] Chaitanya K. Joshi, Arian Rokkum Jamasb, Ramon Viñas Torné, Charles Harris, Simon V Mathis, Alex Morehead, Rishabh Anand, and Pietro Lio. gRNAde: Geometric deep learning for 3D RNA inverse design. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=lvw3UgeVxS.

[6] Dongyue Hou, Shuai Zhang, Mengyao Ma, Hanbo Lin, Zheng Wan, Hui Zhao, Ruian Zhou, Xiao He, Xian Wei, Dianwen Ju, et al. A hyperbolic discrete diffusion 3D RNA inverse folding model for functional RNA design. *Journal of Chemical Information and Modeling*, 65(13): 6568–6584, 2025.

[7] Felix Wong, Dongchen He, Aarti Krishnan, Liang Hong, Alexander Z Wang, Jiuming Wang, Zhihang Hu, Satotaka Omori, Alicia Li, Jiahua Rao, et al. Deep generative design of RNA aptamers using structural predictions. *Nature Computational Science*, 4(11):829–839, 2024.

[8] Laura R Ganser, Megan L Kelly, Daniel Herschlag, and Hashim M Al-Hashimi. The roles of structural dynamics in the cellular functions of RNAs. *Nature Reviews Molecular Cell Biology*, 20(8):474–489, 2019.

[9] Hao Li, Robert Helling, Chao Tang, and Ned Wingreen. Emergence of preferred structures in a simple model of protein folding. *Science*, 273(5275):666–669, 1996.

[10] J. L. England, B. E. Shakhnovich, and E. I. Shakhnovich. Natural selection of more designable folds: A mechanism for thermophilic adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, 100(15):8727–8731, 2003. doi: 10.1073/pnas.1530713100.

[11] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: a case study in RNA secondary structures. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 255(1344):279–284, mar 1994. doi: 10.1098/rspb.1994.0040.

[12] Y. Zhang and J. 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.

[13] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[14] Clement Bernard, Guillaume Postic, Sahar Ghannay, and Fariza Tahi. RNAdvisor: a comprehensive benchmarking tool for the measure and prediction of RNA structural model quality. *Briefings in Bioinformatics*, 25(2), 2024.

[15] Tianshuo Zhou, Ning Dai, Sizhen Li, Max Ward, David H Mathews, and Liang Huang. RNA design via structure-aware multifrontier ensemble optimization. *Bioinformatics*, 39 (Supplement_1):i563–i571, 2023.

[16] Yang Li, Chengxin Zhang, Chenjie Feng, Robin Pearce, P Lydia Freddolino, and Yang Zhang. Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction. *Nature Communications*, 14(1):5745, 2023.

[17] Robin Pearce, Gilbert S Omenn, and Yang Zhang. De novo RNA tertiary structure prediction at atomic resolution using geometric potentials from deep learning. *bioRxiv*, pages 2022–05, 2022.

[18] Yuki Kagaya, Zicong Zhang, Nabil Ibtehaz, Xiao Wang, Tsukasa Nakamura, Pranav Deep Punuru, and Daisuke Kihara. NuFold: end-to-end approach for RNA tertiary structure prediction with flexible nucleobase center representation. *Nature Communications*, 16(1):881, 2025.

[19] Wenkai Wang, Chenjie Feng, Renmin Han, Ziyi Wang, Lisha Ye, Zongyang Du, Hong Wei, Fa Zhang, Zhenling Peng, and Jianyi Yang. trRosettaRNA: automated prediction of RNA 3D structure with transformer network. *Nature Communications*, 14(1):7266, 2023.

[20] Tao Shen, Zhihang Hu, Siqi Sun, Di Liu, Felix Wong, Jiuming Wang, Jiayang Chen, Yixuan Wang, Liang Hong, Jin Xiao, et al. Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nature Methods*, 21(12):2287–2298, 2024.

[21] Minkyung Baek, Ryan McHugh, Ivan Anishchenko, Hanlun Jiang, David Baker, and Frank DiMaio. Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nature Methods*, 21(1):117–121, 2024.

[22] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, 2024.

[23] 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(1):26, 2011.

[24] 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(10):1234–1242, 2022.

[25] 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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RemfXx7ebP.

[26] Cheng Tan, Yijie Zhang, Zhangyang Gao, Hanqun Cao, Siyuan Li, Siqi Ma, Mathieu Blanchette, and Stan Z Li. R3Design: deep tertiary structure-based RNA sequence design and beyond. *Briefings in Bioinformatics*, 26(1):bbae682, 2025.

[27] Ke Liu, Shuaike Shen, and Hao Chen. From sentences to sequences: Rethinking languages in biological system. *arXiv preprint arXiv:2507.00953*, 2025.

[28] Sharat Patil, Frederic Runge, Jörg KH Franke, and Frank Hutter. Towards generative RNA design with tertiary interactions. In *ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2024.

[29] Shengfan Wang, Hong-Bin Shen, and Xiaoyong Pan. Alignment-driven cross-graph modeling for 3D RNA inverse folding. *bioRxiv*, pages 2025–05, 2025.

[30] 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.

[31] Andrew Kubaney, Andrew Favor, Lilian McHugh, Raktim Mitra, Robert Pecoraro, Justas Dauparas, Cameron Glasscock, and David Baker. RNA sequence design and protein–DNA specificity prediction with NA-MPNN. *bioRxiv*, 2025. doi: 10.1101/2025.10.03.679414. URL https://www.biorxiv.org/content/early/2025/10/08/2025.10.03.679414.

[32] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022.

[33] 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*, 2025.

[34] Fanglei Xue, Andrew Kubaney, Zhichun Guo, Joseph K Min, Ge Liu, Yi Yang, and David Baker. Improving protein sequence design through designability preference optimization. *arXiv preprint arXiv:2506.00297*, 2025.

[35] Hanqun Cao, Hongrui Zhang, Junde Xu, Zhou Zhang, Lingdong Shen, Minghao Sun, Ge Liu, Jinbo Xu, Wu-Jun Li, Jinren Ni, et al. From supervision to exploration: What does protein language model learn during reinforcement learning? *arXiv preprint arXiv:2510.01571*, 2025.

[36] Ryan Park, Darren J Hsu, C Brian Roland, Maria Korshunova, Chen Tessler, Shie Mannor, Olivia Viessmann, and Bruno Trentini. Improving inverse folding for peptide design with diversity-regularized direct preference optimization. *arXiv preprint arXiv:2410.19471*, 2024.

[37] 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.

[38] 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. *arXiv preprint arXiv:2410.13643*, 2024.

[39] Mengdi Liu, Xiaoxue Cheng, Zhangyang Gao, Hong Chang, Cheng Tan, Shiguang Shan, and Xilin Chen. ProtInvTree: Deliberate protein inverse folding with reward-guided tree search. *arXiv preprint arXiv:2506.00925*, 2025.

[40] Hanqun Cao, Haosen Shi, Chenyu Wang, Sinno Jialin Pan, and Pheng-Ann Heng. GLID^{2}E: A gradient-free lightweight fine-tune approach for discrete sequence design. In *ICLR 2025 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2025.

[41] Dingyi Rong, Haotian Lu, Wenzhuo Zheng, Fan Zhang, Shuangjia Zheng, and Ning Liu. EnerBridge-DPO: Energy-guided protein inverse folding with Markov bridges and direct preference optimization. *arXiv preprint arXiv:2506.09496*, 2025.

[42] Junqi Liu, Xiaoyang Hou, Chence Shi, Xin Liu, Zhi Yang, and Jian Tang. Property-driven protein inverse folding with multi-objective preference alignment. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=m826DekCpp.

[43] Zhanhui Zhou, Jie Liu, Chao Yang, Jing Shao, Yu Liu, Xiangyu Yue, Wanli Ouyang, and Yu Qiao. Beyond one-preference-for-all: Multi-objective direct preference optimization for language models. *arXiv preprint arXiv:2310.03708*, 2023.

[44] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[45] Yu Pan, Zhongze Cai, Guanting Chen, Huaiyang Zhong, and Chonghuan Wang. What matters in data for DPO? *arXiv preprint arXiv:2508.18312*, 2025.

[46] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[47] Marc Parisien, José Almeida Cruz, Eric Westhof, and François Major. New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, 15(10): 1875–1885, 2009.

[48] Ian W Davis, Andrew Leaver-Fay, Vincent B Chen, Jeremy N Block, Gary J Kapral, Xueyi Wang, Laura W Murray, W Bryan Arendall III, Jack Snoeyink, Jane S Richardson, et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Research*, 35(suppl_2):W375–W383, 2007.

[49] Rhiju Das, John Karanicolas, and David Baker. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature Methods*, 7(4):291–294, 2010. doi: 10.1038/nmeth.1433.

[50] Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707. URL https://www.biorxiv.org/content/10.1101/2025.06.14.659707.

[51] ByteDance AML AI4Science Team, Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, Shenghao Wu, Kuangqi Zhou, Yanping Yang, Zhenyu Liu, Lan Wang, Bo Shi, Shaochen Shi, and Wenzhi Xiao. Protenix - advancing structure prediction through a comprehensive AlphaFold3 reproduction. *bioRxiv*, 2025. doi: 10.1101/2025.01.08.631967. URL https://www.biorxiv.org/content/10.1101/2025.01.08.631967.

[52] Peter Kerpedjiev, Stefan Hammer, and Ivo L Hofacker. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics*, 31(20):3377–3379, 2015.

[53] Schrödinger, LLC and Warren DeLano. PyMOL. Version 2.4.0, 2020. URL http://www.pymol.org/pymol.

[54] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

[55] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=iUwHnoENnl.

[56] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

[57] Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[58] Mark E Fornace, Jining Huang, Cody T Newman, Avinash Nanjundiah, Nicholas J Porubsky, Marshall B Pierce, and Niles A Pierce. NUPACK: Computational nucleic acid analysis and design. *ACS Synthetic Biology*, 2026.

[59] Stanislav Bellaousov, Jessica S Reuter, Matthew G Seetin, and David H Mathews. RNAstructure: web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Research*, 41 (W1):W471–W474, 2013.

[60] Yang Li, Chenjie Feng, Xi Zhang, Sho Tsukiyama, Duanyu Feng, and Yang Zhang. DRfold2 is a deep learning-based tool that enables efficient and accurate RNA structure prediction. *PLoS Biology*, 24(2):e3003659, 2026.

[61] Chengxin Zhang, Morgan Shine, Anna Marie Pyle, and Yang Zhang. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature Methods*, 19(9):1105–1108, 2022. doi: 10.1038/s41592-022-01575-6.

Appendix

A Variability-aware margins and candidate-pool coverage

Standardized margins. We set $\varepsilon_{r,m} = \alpha_r \sigma_m$, where σ_m is the standard deviation of proxy m over the candidate pool. Retaining a pair requires $[\phi_m(s^w) - \phi_m(s^l)]/\sigma_m > \alpha_r$ for every metric. Thus a positive rescaling of a proxy and its standard deviation leaves the rule unchanged. The schedule $\alpha_{1,2} = 0.25$ and $\alpha_{3,4,5} = 0.125$ selects well-separated pairs early and finer distinctions later.

Gaussian comparison motivation. For one metric, suppose the observed difference is $\widehat{\Delta}_m = \Delta_m^* + \eta_m$, with $\eta_m \sim \mathcal{N}(0, 2\tau_m^2)$ with $\tau_m > 0$, arising from independent, equal-variance errors on the two candidates. Then

$$\Pr(\widehat{\Delta}_m > 0 \mid \Delta_m^*) = \Phi\left(\frac{\Delta_m^*}{\sqrt{2}\tau_m}\right).$$

This probit comparison illustrates why the meaning of a raw margin depends on the proxy's scale. The implemented rule uses candidate-pool standard deviations σ_m as empirical scale parameters and applies the standardized threshold coordinatewise.

Label-level scalarization consistency. If $\phi_m(s^w) - \phi_m(s^l) > \varepsilon_{r,m}$ for every m, then for any $w \in \Delta^{M-1}$,

$$w^\top \phi(s^w) > w^\top \phi(s^l) + w^\top \varepsilon_r.$$

The retained labels therefore agree with every nonnegative scalarization with nonzero total weight. The structural gate further requires each winner to meet the specified pLDDT and RMSD thresholds.

Empirical pair composition. Among 54,540 evaluated ordered candidate pairs from the training pool, 22.3% are ε-Pareto-strict, 21.5% improve MFE while degrading pLDDT or RMSD, 20.9% have mixed signs on the structural axes, and 35.3% are dominated by the loser. The MFE-only improvement group illustrates the pairs rejected by the joint criterion but potentially accepted by an MFE-dominated scalarization.

Pair-set coverage. At $\varepsilon_{r,m} = 0.125\sigma_m$, the training set contains 20,811 pairs, with 7,591 distinct winner sequences across 2,071 backbones and approximately 8,721 distinct winner–backbone tuples. Evaluation uncertainty is aggregated across test targets (App. I).

B Fixed-reference variational characterization

Setup. Fix a backbone $\mathcal{G}$ and suppress it in the notation. Let $q(s) > 0$ on the finite sequence space, let u be a fixed reward with $u_{\min} \le u(s) \le u_{\max}$, and let $\beta > 0$. The KL-regularized objective underlying DPO [44] is

$$J_q(\pi; u) = \mathbb{E}_\pi[u] - \beta \mathrm{KL}(\pi \| q).$$

Set $Z = \sum_s q(s) e^{u(s)/\beta}$ and $\pi^*(s) = q(s) e^{u(s)/\beta}/Z$.

Proof of Proposition 1. Since $\log(\pi^*/q) = u/\beta - \log Z$, for every policy π,

$$\begin{aligned}\beta \mathrm{KL}(\pi \| \pi^*) &= \beta \mathrm{KL}(\pi \| q) - \mathbb{E}_\pi[u] + \beta \log Z \\ &= \beta \log Z - J_q(\pi; u).\end{aligned}$$

At $\pi = \pi^*$ this gives $J_q(\pi^*; u) = \beta \log Z$, and hence $J_q(\pi^*; u) - J_q(\pi; u) = \beta \mathrm{KL}(\pi \| \pi^*)$. Nonnegativity of KL establishes that π^* is the unique maximizer. Moreover, $\log Z \ge u_{\min}/\beta$, so

$$\mathrm{KL}(\pi^* \| q) = \frac{\mathbb{E}_{\pi^*}[u]}{\beta} - \log Z \le \frac{u_{\max} - u_{\min}}{\beta}.$$

□

Role in RiboPO. This standard identity characterizes the optimum of a fixed-reward objective relative to its reference. Keeping $q = \pi_{\mathrm{ref}}$ unchanged gives each round the same reference distribution. The effect of this design on the implemented optimizer is assessed by the multi-seed frozen-reference ablation (App. H).

C Pair-distribution shift and importance reweighting

Setup. At a fixed round and parameter θ, let $P = \nu_{\text{ref}}$ and $Q = \nu_\theta$ be the distributions of preference pairs formed from reference-policy and current-policy candidates under the same selection rule. For a pair $z = (s_w, s_l, \mathcal{G})$, let $g_\theta(z) = \nabla_\theta \ell_{\text{DPO}}(\theta; z)$ and define $\Gamma(\theta; \nu) = \mathbb{E}_{z\sim\nu}[g_\theta(z)]$. The sampling distribution is held fixed when taking this per-pair derivative. Assume $Q \ll P$, $\|g_\theta(z)\|_2 \le G$, and a finite second moment for $w(z) = dQ/dP(z)$. Define the Rényi-2 divergence by

$$D_2(Q\|P) = \log \mathbb{E}_P[w^2].$$

These are distributions after preference selection, so the pair formulation includes dependence induced by the selection rule.

Proof of Proposition 2. Change of measure gives $\Gamma(\theta; Q) - \Gamma(\theta; P) = \mathbb{E}_P[(w-1)g_\theta]$. Since $\mathbb{E}_P[w] = 1$, Cauchy–Schwarz yields

$$\begin{aligned}\|\Gamma(\theta; Q) - \Gamma(\theta; P)\|_2 &\le \sqrt{\mathbb{E}_P[(w-1)^2]}\,\sqrt{\mathbb{E}_P[\|g_\theta\|_2^2]} \\ &\le G\sqrt{\mathbb{E}_P[w^2] - 1} = G\sqrt{e^{D_2(Q\|P)} - 1}.\end{aligned}$$

□ This expected-update bound quantifies the dependence on pair-distribution shift. Exact importance weighting satisfies $\mathbb{E}_P[w g_\theta] = \Gamma(\theta; Q)$; large second moments motivate controlling the variability of the weights [56, 57].

Implemented weighting rule. For rounds $r \ge 2$, we temper the product of the two sequence likelihood ratios by a geometric mean and clip it:

$$w_{\text{pair}}^{(r)} \;=\; \min\Big(\bar\rho,\; \sqrt{\tfrac{\pi_{\theta,r-1}(s_w)\,\pi_{\theta,r-1}(s_l)}{\pi_{\text{ref}}(s_w)\,\pi_{\text{ref}}(s_l)}}\Big), \qquad \bar\rho{=}5. \tag{3}$$

All policy probabilities condition on the same backbone. We then normalize the weights to unit mean within each batch and use them to weight the DPO loss. The unnormalized weights are evaluated at the preceding-round checkpoint and fixed within a round; normalization is applied separately in each batch. Proposition 2 concerns the exact pair-distribution ratio; Eq. 3 is the practical tempered estimator used for variance control. Extended-round comparisons and weight diagnostics are reported in §4.6 and App. L.

D Stage-2 architecture and training

FiLM head. The Stage-2 wrapper modulates the gRNAde decoder's scalar features as $h \to h(1 + \gamma(w)) + \beta(w)$, where $\gamma, \beta : \mathbb{R}^M \to \mathbb{R}^{h_{\text{dim}}}$ are linear layers initialized to zero. Total: $M \cdot h_{\text{dim}} \cdot 2 + 2h_{\text{dim}} = 1024$ parameters. The residual init guarantees $\pi_w(\cdot)|_{w\text{ unset}} \equiv \pi_{\text{base}}(\cdot)$ at initialization.

Training. Per batch, we sample $w \sim \text{Dirichlet}(1)$ (uniform on the simplex). The DPO loss is computed on the ε-Pareto-dominant pair set; the policy's log-prob conditions on the sampled w.

Inference. Querying $w \in \Delta^{M-1}$ produces conditional samples; we evaluate the centroid and three simplex vertices. The default is the centroid $w = (1/M, \ldots, 1/M)$.

E Full vanilla-baseline numbers (with leakage marker)

Table 4 reports full-set results for four vanilla baselines. Table 2 reports the corresponding baseline-specific overlap-audited comparisons.

Baseline-specific overlap audits. The overlap-audited subsets exclude training overlap separately for each external method. The audit uses released training splits at PDB-identifier level for RDesign/R3Design and RhoDesign, and the released 80% sequence-identity criterion for RiboDiffusion. Because the resulting targets differ by method, Table 2 provides descriptive results for these subsets.

F Evaluation pipeline and reproducibility

Evaluation pipeline. SSTT scoring runs ViennaRNA's partition-function ensemble on each (target dot-bracket, designed sequence) pair to compute MFE, ensemble defect, $P(S_0)$, and T_m. Length-mismatched dot-brackets (∼1.2% of multi-chain test structures, primarily multi-chain inputs) are

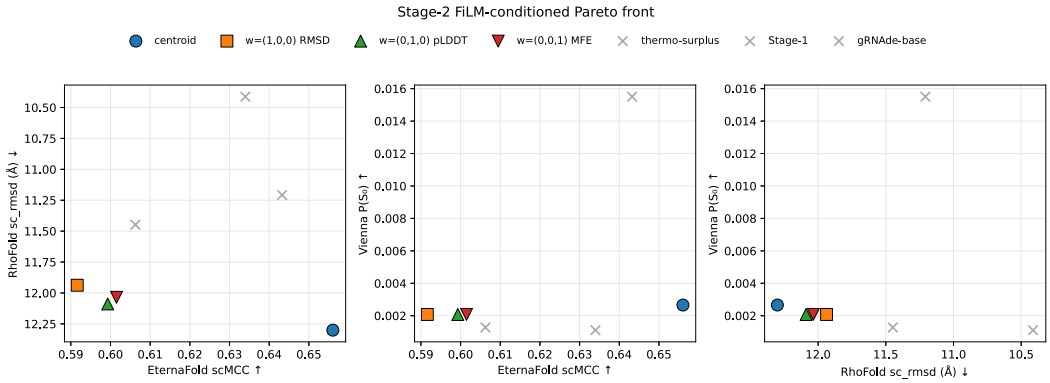


Figure 5: **Stage-2 FiLM-conditioned policy: 2-mode behavior.** Stage-2 trained checkpoint sampled at four w vectors (centroid + 3 simplex vertices). Three projections of the (scMCC, RMSD, $P(S_0)$) space. Centroid ($\circ$) sits at high scMCC; the three corner-w queries ($\square$, $\triangle$, ∇) converge to a single thermo-strong mode. Stage-1, thermo-surplus, and gRNAde-base shown for context.

Table 4: Full-set vanilla baselines on DAS test (no preference optimization). Rows marked † have substantial overlap between training data and the DAS test split; retained for descriptive context. The direct primary comparison uses the shared gRNAde–R2 test set in Table 1.

Model	Rec.	scMCC	RMSD	TM	MFE
RhoDesign†	0.543	0.20	15.13	0.204	-21.93
RIdiffusion	0.412	0.26	15.98	0.129	-24.34
NA-MPNN	0.459	0.30	11.43	0.271	-28.42
RiboDiffusion†	0.565	0.33	9.55	0.338	-23.86

soft-truncated to the designed sequence length and rebalanced to valid dot-bracket notation before ViennaRNA evaluation. Per-structure means exclude missing or invalid (NaN) outputs. Tables 1–2 use this pipeline. The ablation and cross-engine panels use separate evaluation snapshots, as specified in their captions. The evaluator and its regression tests are released with the code.

Split and optimization details. We use the gRNAde v0.3.2 single-state split, constructed by US-align/qTMclust at a structural-similarity cutoff of 0.45. The 14 clusters containing the Das et al. targets supply 98 test entries; remaining one- or two-entry clusters fill the 100-entry validation set, with 4,025 entries assigned to training. Stored train/test structural-cluster IDs are disjoint. The cutoff is a clustering parameter. CD-HIT-EST preprocessing additionally groups sequences at 80% identity. The reference checkpoint is `gRNAde_ARv1_1state_das.h5`. The primary loss uses $\beta = 0.12$, SFT weight $\lambda_{\text{SFT}} = 0.10$, and learning rate 1.8×10^{-4}. Margin and gate parameters are defined in §3.2; the replication schedules are separated in Appendix H.

G Multi-round design ablation

Table 5 compares the component-ablation configurations. Reference updating favors lower MFE but gives worse RMSD, and the no-gate control shifts toward lower MFE with reduced scMCC in this run. The constant-margin and decreasing-margin configurations differ more modestly. The independent-training-seed replications in Appendix H assess the reproducibility of these observations and distinguish compositional behavior from seed-to-seed stability.

H Independent-training-seed replications

We evaluated independently trained checkpoints at 8 designs per target, sampling temperature 0.1, and evaluation seed 42, aggregating designs within each target before aggregation across targets. The gate and reference comparisons use a compute-reduced schedule of four preference-optimization

Table 5: Design ablations on DAS. This panel uses a separate evaluation snapshot, so absolute MFE values differ from the main-table results. Comparisons are within this table; shading identifies the full configuration. Red highlights the reward-hacking diagnostics discussed in the text. Independent-training-seed replications are reported in Appendix H.

Variant	**Rec.** ↑	**scMCC** ↑	**RMSD** ↓	**TM** ↑	**MFE** ↓
gRNAde (base)	0.529	0.60	11.45	0.29	-32.83
DPO (single round, no curriculum)	0.504	0.71	10.40	0.30	-35.83
+ **Reference Model Update** (per-round)	0.482	0.66	14.30	0.16	-48.07
+ Constant margin α_r=0.25	0.508	0.70	10.55	0.28	-34.91
+ No structural quality gate	0.499	0.64	11.81	0.27	-39.42
+ **Frozen ref + decreasing margin** (ours)	0.506	0.72	10.42	0.29	-35.22

rounds with three epochs per round. Their absolute scores are therefore reported separately from the primary R2 configuration. These replications assess recovery, EternaFold scMCC, GC content, and ViennaRNA MFE; tertiary predictors were not re-run across all replicated arms. Unless stated otherwise, the dispersion reported below is the standard deviation across training seeds.

MFE-only/no-gate control. The MFE-only/no-gate control increases GC content and lowers MFE at each of the three matched seeds. Table 6 separates the available arm summaries from the matched-seed contrasts: the control has three seeds, whereas the gated/frozen-reference summary has five. The compositional signature is clearer than the changes in recovery and scMCC.

Table 6: R4 endpoints for the MFE-only/no-gate control and gated/frozen-reference runs. Upper panel: mean ± SD over each arm's available seeds. Lower panel: matched-seed control-minus-gated differences. MFE is in kcal/mol.

Arm (number of seeds)	GC	MFE	scMCC	Recovery
MFE-only/no-gate (3)	0.686 ± 0.021	-40.6 ± 1.5	0.625 ± 0.039	0.486 ± 0.008
Gated/frozen reference (5)	0.624 ± 0.027	-34.6 ± 1.5	0.641 ± 0.024	0.496 ± 0.006
Matched training seed	ΔGC	ΔMFE		
7	+0.045	-6.0		
42	+0.092	-7.9		
123	+0.049	-4.3		

Frozen versus updated reference. The five matched training seeds are 0, 6, 7, 42, and 123. The frozen reference gives tighter seed-to-seed dispersion in recovery and scMCC (Table 7). Reference updating improves scMCC at some seeds, while its lowest-recovery seed reaches 0.434; frozen-reference recovery remains between 0.491 and 0.507. The reproducible advantage is training stability across seeds.

Table 7: Frozen versus updated reference at R4 over five matched seeds. Values are mean ± SD; MFE is in kcal/mol.

Arm	Recovery	scMCC	MFE
Frozen reference	0.496 ± 0.006	0.641 ± 0.024	-34.6 ± 1.5
Reference update	0.484 ± 0.026	0.615 ± 0.046	-35.6 ± 1.8
SD ratio (updated/frozen)	4.2	1.9	1.2

Decreasing versus constant margins. At the final 8-design re-evaluation, decreasing margins give higher scMCC at both available matched seeds, with equal recovery (Table 8). The table reports these two matched training seeds.

Matched-seed static-pair versus clipped-IS training. Both arms are trained through round 5 at seeds 7 and 123, with the seed-42 trajectory reported in Appendix L. At both seeds, static-pair reuse

Table 8: EternaFold scMCC for decreasing and constant margins at the final 8-design re-evaluation. The last row gives mean ± SD.

Training seed	Decreasing margin	Constant margin
7	0.654	0.631
123	0.677	0.661
Mean ± SD	0.666 ± 0.012	0.646 ± 0.015

lowers MFE and reduces recovery over rounds, whereas clipped-IS preserves or slightly improves recovery (Table 9). GC still increases in the clipped-IS runs; recovery preservation is the consistent replicated distinction.

Table 9: Matched-seed static-pair versus clipped-IS comparisons. Entries show R1 → R5; MFE is in kcal/mol. These trajectories extend beyond the four-round gate/reference experiments.

Training seed	Arm	GC	MFE	Recovery
7	Static pairs	0.581 → 0.622	-32.3 → -35.7	0.497 → 0.466
7	Clipped-IS	0.566 → 0.619	-31.4 → -30.5	0.483 → 0.499
123	Static pairs	0.559 → 0.605	-30.7 → -33.5	0.523 → 0.485
123	Clipped-IS	0.564 → 0.597	-31.7 → -30.6	0.483 → 0.503

I Statistical summaries and ensemble observations

The statistical analysis uses target-level percentile bootstrap intervals (10,000 resamples) and paired Wilcoxon tests. The independent-predictor tests are tabulated in Appendix O; training-seed variation is reported separately in Appendix H.

Secondary ensemble observations. For R2, target ensemble defect per nucleotide changes descriptively from 0.0101 to 0.0093. The raw target-probability contrast is +0.0103 with paired p = 0.20; the GC/length-adjusted log-probability contrast is also non-significant. The thermodynamic-surplus checkpoint separately changes $P(S_0)$ from 0.00128 to 0.0155, with a paired p-value below 0.001. These are distinct checkpoint-specific results. Target defect has not been adjusted with a target-pairing-matched composition model, so the ensemble quantities are retained as secondary descriptive observations.

J Single-policy operating points

Beta sweep. We trained DPO at $\beta \in \{0.001, 0.05, 0.12, 1.0\}$ on the α_r=0.25 pair set. At β=0.001, GC reaches 0.69, with RMSD 14.7 Å and scMCC 0.61; β=1.00 gives GC 0.70 and RMSD 12.7 Å. The chosen β=0.12 yields a more balanced operating point (recovery 0.50, scMCC 0.69, RMSD 10.2 Å, GC 0.59). These comparisons indicate sensitivity to the preference scale. Per-configuration numbers are released with the supplementary code.

Loss ablation. On the same pair set and training budget, IPO [54], KTO [55], and Pareto-DPO Stage 1 (stochastic-w Dirichlet sampling, no FiLM) each underperform standard DPO. Among the alternative losses, KTO has the highest scMCC (0.66) but worst RMSD (14.2 Å), supporting the use of standard DPO in this setting.

The Stage-1 and Stage-2 operating points are reported in Table 1. The FiLM architecture and centroid/vertex queries are described in Appendix D.

K Importance-corrected iterative DPO

The implementation uses Eq. 3: sequence likelihood ratios from the preceding-round checkpoint are combined through a geometric mean, clipped at 5, and normalized to unit batch mean. The resulting

batch-normalized weights multiply the DPO loss at rounds $r \geq 2$. Appendix C distinguishes this practical weighting rule from the exact pair-distribution ratio. Five-round trajectories are reported in Table 10, and additional matched training seeds are reported in Appendix H.

L Extended rounds, anchored log-prob L^1 deviation

Per-round trajectories of the two R=5 multi-round trainings are summarized in Table 10; the released supplementary reports the same data per-structure with bootstrap CIs.

Table 10: Per-round metrics for the two R=5 multi-round trainings: *static-pair (off)* (no importance correction) vs. *clipped-IS (on)* (importance correction at $\bar{\rho}$=5). Both share the same β=0.12, λ_{SFT}=0.10 and frozen-reference DPO loss; they differ only in importance reweighting at rounds $r \geq 2$. These matched-seed trajectories show stronger late-round structural success with clipped reweighting.

Variant	R	scMCC	$P(S_0)$	pLDDT≥0.7	RMSD≤8Å	T_m (°C)	GC
static-pair (off)	1	0.7233	0.01565	0.364	0.481	26.64	0.583
	2	0.6349	0.00182	0.369	0.417	28.67	0.598
	3	0.6570	0.00310	0.346	0.372	28.96	0.596
	4	0.6785	0.00210	0.421	0.364	25.39	0.608
	5	0.6696	0.00197	0.332	0.246	29.73	0.632
clipped-IS (on)	1	0.7108	0.00142	0.427	0.459	27.11	0.563
	2	0.6551	0.00793	0.356	0.406	26.93	0.584
	3	0.6788	0.01120	0.438	0.472	26.86	0.611
	4	0.6828	0.00430	0.378	0.393	24.61	0.577
	5	0.6615	0.00224	0.454	0.486	27.44	0.598

In the static-pair run, RMSD success decreases from 0.481 to 0.246 between rounds 1 and 5, while GC rises overall from 0.583 to 0.632. The clipped-IS run ends with RMSD success 0.486 and pLDDT success 0.454. Its GC varies between 0.563 and 0.611 across rounds. These trajectories illustrate the behavior of the two runs; the additional matched-seed results in Appendix H identify recovery retention as the consistent replicated benefit.

Anchored log-prob L^1 deviation across rounds. We report the mean absolute change in teacher-forced log-likelihood on 50 held-out preference pairs, averaging over winners and losers:

$$\widehat{|\Delta_R|} = \mathbb{E}_{(s,\mathcal{G})\sim\mathcal{D}_{\text{test}}}\left[\left|\log \pi_{\theta_R}(s \mid \mathcal{G}) - \log \pi_{\text{ref}}(s \mid \mathcal{G})\right|\right].$$

This is a fixed-pair likelihood diagnostic, rather than a sampled-policy KL estimate. The static-pair values are 64.6, 70.4, 76.7, 85.5, and 93.0 nats across rounds 1–5; clipped-IS gives 62.7, 21.4, 31.0, 29.6, and 18.7 nats. The latter run remains closer to the reference in this diagnostic after reweighting begins. These measurements describe likelihood changes on the fixed evaluation pairs and complement the structural metrics in Table 10.

Importance-weight ESS and clipping diagnostics. On a sample of 500 train pairs we compute, at each round $r \geq 2$, the un-normalized geometric-mean pair ratio actually used by the trainer, $w_{\text{raw}}^{(r)} = \sqrt{\pi_{\theta,r-1}(s_w)\pi_{\theta,r-1}(s_l)/\pi_{\text{ref}}(s_w)\pi_{\text{ref}}(s_l)}$, the upper-clip fraction at $\bar{\rho}$=5, and effective sample size ESS$=(\sum_i w_i)^2/\sum_i w_i^2$ using clipped weights $w_i = \min(5, w_{\text{raw},i}^{(r)})$. Table 11 reports the clipped-IS run. Within a batch, ESS is invariant to a common normalization factor. The pair ratio is heavy-tailed throughout (median $\approx$ 0, the bulk of pairs has $\pi_{\theta,r-1} \ll \pi_{\text{ref}}$), but ESS recovers from 0.4% at R=2 to 24–35% at $R \geq 3$, after reweighting begins; concurrently the upper-clip fraction grows from 0.4% to 28.6%, showing that clipping controls an increasing fraction of the high-ratio pairs.

M Matched-utility reranking

We apply the same Pareto-feasible-then-max-$P(S_0)$ selector to each model's candidate pool on 98 DAS targets. Table 12 compares candidate budgets under this shared selection rule. The RIBOPO

Table 11: IS diagnostics on 500 training pairs. The second column reports mean log-ratio; the upper-clip fraction uses the unnormalized geometric-mean ratio and threshold 5. ESS is reported for clipped weights.

Round	Mean log-ratio (nats)	Upper-clip fraction	ESS/N
2	-66.3	0.4%	0.4%
3	-12.6	20.8%	24.3%
4	-14.8	22.4%	26.4%
5	-11.6	28.6%	34.6%

rows use the R2 checkpoints of the dedicated static-pair and clipped-IS experiments, separate from the default checkpoint in Table 1.

Table 12: Matched-utility reranking on DAS. C is the number of candidates per target; threshold columns report success fractions. Both RIBOPO extension runs match gRNAde's best-of-64 joint success with eight candidates.

Method	C	scMCC	pLDDT$\geq$0.7	RMSD$\leq$8Å	pass@1
gRNAde	8	0.616	0.459	0.439	0.388
gRNAde	64	0.640	0.469	0.469	0.408
Static-pair DPO (R2)	8	0.618	0.439	0.459	0.408
Clipped-IS DPO (R2)	8	0.623	0.459	0.480	0.408

N Cross-engine validation

To assess whether the thermodynamic gains generalize beyond the training oracle, we re-scored the multi-round R2 designs with two thermodynamic engines orthogonal to ViennaRNA: NUPACK [58] and RNAstructure [59].

Table 13: Cross-engine MFE robustness. RiboPO's gain over gRNAde-base is consistent across three independent engines, supporting robustness of the energy-score difference across engines. The paired cross-engine panel uses its own evaluation snapshot; comparisons are within each column.

Method	**Vienna MFE**	**NUPACK MFE**	**RNAstructure MFE**
gRNAde (Base)	-30.37	-28.90	-30.53
RiboPO (R2)	-33.95	-31.84	-34.29
Δ	$\mathbf{-11.8\%}$	$\mathbf{-10.2\%}$	$\mathbf{-12.3\%}$

O Independent 3D structure predictors

We evaluated the same fixed RiboPO-R2 and gRNAde sequences with Boltz-2, AlphaFold3 (AF3), and Protenix, outside the RhoFold+ preference-construction loop. Predictor outputs were not used for sequence filtering or reranking. Table 14 reports complete target-paired results; Boltz-2 and AF3 cover 98 targets, and Protenix covers 93 after excluding five targets with incomplete method pairs.

Inference protocols. Boltz-2 v2.2.1 used default weights, single-sequence/no-MSA mode, one diffusion sample, and 20 sampling steps. AF3 used a query-only alignment (depth 1 or MSA-free) and no target templates. Protenix package v0.5.5 used checkpoint `protenix_base_default_v0.5.0` with `use_msa=False`. All three predictors used seed 42. Single-sequence inference keeps homolog retrieval and MSA depth from varying with the designed sequence. The paired tests quantify variation across targets under these fixed inference settings.

A separate 16-target panel evaluated AF3 and DRFold2. AF3 RMSD decreased from 3.00 to 2.81 Å and TM-score increased from 0.41 to 0.43 for gRNAde versus R2; DRFold2 RMSD decreased by 5.7%. The full-DAS panels above provide the primary cross-evaluator comparison.

Table 14: Independent-predictor evaluation on DAS. Each cell reports gRNAde → RiboPO-R2, followed by the two-sided paired Wilcoxon p-value. N is the number of complete target pairs; all inference uses seed 42.

Predictor	N	RMSD (Å)	TM-score	pLDDT
Boltz-2	98	3.256 → 3.178 p = 0.055	0.398 → 0.411 p = 0.042	0.624 → 0.656 p = 0.025
AF3	98	3.286 → 3.157 p = 0.074	0.372 → 0.375 p = 0.743	0.576 → 0.587 p = 0.375
Protenix-base	93	3.186 → 3.029 p = 0.044	0.399 → 0.414 p = 0.031	0.638 → 0.649 p = 0.628

P Thermodynamic-surplus filter detail

The thermodynamic-surplus variant uses a GC-controlled pair set: we fit MFE on length and length-scaled GC content over the candidate pool and retain pairs whose winner strictly improves the GC-corrected residual. This removes 23.7% of pairs at the α=0.25 margin. The resulting policy improves target-structure probability and melting temperature alongside secondary-structure consistency (Table 15). Vienna MFE is -30.74 kcal/mol, a 7.3% improvement over gRNAde-base. The full per-structure table is released with the supplementary code.

Scope of the regression. The fitted model adjusts for length and length-scaled GC content. In the R2 sensitivity analysis, fitting on 784 reference designs or pooling all 1,568 designs retains 69.4% or 64.9% of the raw MFE contrast, respectively. These percentages use the matched regression panel's raw contrast of -3.57 kcal/mol; the main-table aggregates describe a separate evaluation snapshot. The retained residual describes the contrast after the specified adjustment. Target-structure probability and ensemble defect require their own composition analyses and are reported separately in Appendix I.

Table 15: Thermodynamic-surplus filter detail: per-axis SSTT breakdown vs. gRNAde-base (full per-structure table released with the supplementary code).

Axis	**gRNAde-base**	**RIBOPO thermo-surplus**
scMCC	0.606	0.643
RMSD (Å)	11.45	11.21
pLDDT≥0.7	0.381	0.438
RMSD≤8Å	0.416	0.430
$P(S_0)$	0.00128	0.0155
T_m (°C)	27.0	28.4
Recovery	0.529	0.496

Q Structural-gate (κ, γ) sensitivity surface

We re-construct preference-pair sets at $(\kappa, \gamma) \in \{0.60, 0.65, 0.70, 0.75, 0.80\} \times \{6, 8, 10, 12\}$ on the cached candidate-pool scores (no retraining required). Table 16 reports the resulting pair-yield surface within the 12,137 ε-Pareto-strict candidate pairs. The chosen operating point $(\kappa, \gamma) = (0.70, 8.0)$ retains 10,845 pairs (89.4% of the ε-Pareto-strict pool); loosening to $(\kappa, \gamma) = (0.60, 10)$ adds only 4.7% pairs, and tightening to $\gamma = 6$ drops 8.1%. Tightening κ to 0.80 drops yield by 58%, indicating a sharp threshold around the RhoFold+ pLDDT distribution upper quartile. The chosen point is at the knee of the strictness-yield curve.

Mean winner GC content across the same grid varies little (0.598–0.606), indicating that changing the gate thresholds has limited effect on winner GC within this Pareto-filtered candidate pool. The thermodynamic-surplus filter separately selects improvements after GC/length adjustment (§3.2).

Table 16: Gate-sensitivity surface on candidate pool: *Left:* pair count surviving ε-Pareto-strictness + (κ, γ) structural gate. *Right:* mean winner GC content under the same gate. Operating point $(0.70, 8.0)$ in **bold**.

	Surviving pairs				**Mean winner GC**			
$\kappa \backslash \gamma$	6	8	10	12	6	8	10	12
0.60	10137	11164	11356	11492	0.606	0.603	0.603	0.602
0.65	10083	11041	11180	11296	0.606	0.604	0.603	0.603
0.70	9965	**10845**	10953	11051	0.606	**0.604**	0.604	0.603
0.75	7598	7995	8030	8064	0.604	0.603	0.603	0.603
0.80	4485	4572	4572	4572	0.598	0.598	0.598	0.598

R Loss-component ablation (DPO vs. SFT)

To isolate the contribution of each loss term, we ablate $\mathcal{L}_{\mathrm{DPO}}$ and $\mathcal{L}_{\mathrm{SFT}}$ in a single-round run with otherwise matched hyperparameters (Table 17). Removing $\mathcal{L}_{\mathrm{DPO}}$ gives an SFT-on-Pareto-winners baseline, with scMCC near the gRNAde-base level and little improvement in tertiary metrics. This comparison supports learning pairwise preferences beyond imitating selected winners. Removing $\mathcal{L}_{\mathrm{SFT}}$ lowers MFE but worsens RMSD and TM-score. The two terms thus play complementary roles: DPO learns from the preference ordering, while the SFT anchor helps maintain structural fidelity during optimization.

Table 17: Loss-component ablation on single-round DPO. Removing $\mathcal{L}_{\mathrm{DPO}}$ gives an SFT-on-winners baseline; removing $\mathcal{L}_{\mathrm{SFT}}$ lowers MFE at the expense of 3D fidelity. Shading identifies the full configuration.

Configuration	**Rec.↑**	**scMCC↑**	**pLDDT↑**	**RMSD (Å)↓**	**TM↑**	**MFE↓**
gRNAde (base, no preference)	0.529	0.61	0.62	11.45	0.29	-28.64
w/o $\mathcal{L}_{\mathrm{DPO}}$ (β=0, SFT-on-winners)	0.490	0.61	0.62	11.50	0.27	-33.68
w/o $\mathcal{L}_{\mathrm{SFT}}$ (λ_{SFT}=0)	0.440	0.64	0.62	13.55	0.19	-42.60
RiboPO (full, β=0.12, λ_{SFT}=0.10)	0.500	0.71	0.66	10.40	0.30	-35.83

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: *The abstract and introduction describe variability-aware Pareto preferences, structural quality control, frozen-reference optimization, and SSTT evaluation centered on RiboPO-R2. Main results and independent-predictor evaluation appear in §4; optional operating points are identified separately.*

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: *The scope is stated in §5: evaluation uses computational structure and thermodynamic proxies, sequence recovery is a distinct objective, and GC/length adjustment addresses the specified compositional effects. Prospective application-specific experiments remain a future validation step.*

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: *Proposition 1 is the standard fixed-reference variational characterization for a fixed bounded reward on a finite sequence space, proved in Appendix B. Proposition 2 bounds the pair-distribution expected-update discrepancy under support, gradient-norm, and second-moment assumptions, proved in Appendix C. The implemented weighting rule and empirical stability evidence are distinguished from these mathematical results.*

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: *Methods specify preference construction, the gate, the loss and the default R2 checkpoint. Appendix F gives split construction and primary optimization settings; Appendices O and H document predictor protocols and replicated training comparisons.*

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: *The full RiboPO codebase, the Vienna evaluation pipeline, the SSTT panel, training and eval configs, preference pair JSONLs, and per-structure result files will be released with the paper. Public DAS benchmark splits are reused from Das et al. [49] as packaged by Joshi et al. [5]; the canonical upstream gRNAde checkpoint is from Joshi et al. [5].*

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: *The main text states the benchmark, default checkpoint, evaluation panel and sampling budget. Appendix F gives the primary loss settings and split procedure; Appendix H separates the reduced-schedule replications, and Appendix O gives independent-predictor settings.*

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: *Target-level statistical methods and secondary ensemble observations are described in Appendix I. Complete-pair counts and paired tests for independent predictors are in Appendix O. Appendix H reports training-seed means, standard deviations and matched-seed contrasts separately from target-level variation.*

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: *The original experiments used one NVIDIA A40 (48 GB) or A100 (80 GB) per run. Original estimates are approximately 6 GPU-h per single-round DPO run, 30 GPU-h per five-round arm, and 3–5 GPU-h per full-SSTT evaluation. The original approximately 200 GPU-h aggregate does not include the added training-seed replications and independent-predictor evaluations; a complete updated compute total is not available.*

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: *The research conforms to the NeurIPS Code of Ethics. The work is computational RNA inverse-folding optimization with no human subjects, no scraped private data, and uses only publicly released RNA structural data and predictor checkpoints.*

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: *The Discussion describes intended RNA-design applications and the need for experimental validation. The paper does not include a separate assessment of negative societal impacts.*

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: *The released artifact is a preference-optimization framework for RNA inverse folding, trained on public structural data. The policy generates short RNA sequences for user-specified backbone geometries.*

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: *Core code/data assets are credited: gRNAde [5] and DAS [49]; baseline/model assets include RhoDesign [7], RIdiffusion [6], RiboDiffusion [30], RDesign [25], R3Design [26], AlignIF [29], and ProteinMPNN/NA-MPNN [32]; evaluation assets include ViennaRNA [23], EternaFold [24], RhoFold+ [20], NUPACK [58], RNAstructure [59], AlphaFold3 [22], DRFold2 [60], Boltz-2 [50], Protenix [51], US-align/TM-score [61, 12], INF [47], and MolProbity [48]. The manuscript credits these assets but does not yet enumerate all license and access terms.*

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: *New assets: (i) the RiboPO training pipeline (DPO with ε-Pareto preference construction, frozen-reference curriculum, optional Stage-2 FiLM head); (ii) the thermodynamic-surplus pair filter; (iii) the Vienna SSTT evaluation pipeline; (iv) the SSTT 21-metric panel definitions; (v) per-structure CIs and effect sizes for the headline claims. All released with documentation in the supplementary code.*

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: *No crowdsourcing or human-subjects research is involved.*

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: *No human-subjects research is involved.*

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: *LLMs were not used as a core component of the methodology or scientific rigor; any usage was limited to writing/editing.*

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.