PROT2RNA: REWARD-GUIDED DISCRETE DIFFUSION FOR PROTEIN-CONDITIONED CDS DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

A single protein can be encoded by many synonymous mRNA coding sequences (CDSs), whose composition affects RNA structure, stability, and protein output. We introduce **Prot2RNA**, a protein-conditioned discrete diffusion model that learns a conditional distribution over natural human CDSs. Motivated by the role of RNA structure in protein output, we shift this conditional distribution toward sequences with increased predicted structural stability using normalized minimum free energy (MFE) as a non-differentiable sequence-level reward, through reinforcement finetuning across diverse proteins (**Prot2RNA-RFT**), followed by test-time reinforcement learning for a fixed target (**Prot2RNA-TTRL**). Prot2RNA-RFT consistently shifts normalized MFE in the intended direction across held-out proteins, while Prot2RNA-TTRL further adapts the policy for a single NanoLuc protein. In HEK293T cells, the selected NanoLuc design achieves the highest measured reporter output among the tested constructs, with 39% and 71% higher output than GEMORNA and LinearDesign, respectively. These results demonstrate that a learned protein-conditioned CDS prior can be effectively optimized toward a sequence-level reward using reinforcement learning, and when adapted to a single target, yield experimentally validated designs that outperform established CDS-design methods.

1 INTRODUCTION

mRNA has emerged as an important therapeutic modality, with applications including vaccines and transient production of therapeutic proteins (Pardi et al., 2018; Sahin et al., 2014). Although the coding sequence (CDS) determines the encoded protein, most amino acids can be represented by multiple synonymous codons. A single protein therefore corresponds to a large space of possible CDSs whose sequence composition can substantially affect RNA structure, transcript stability, translation dynamics, and protein output (Mauger et al., 2019; Leppek et al., 2022; Bicknell et al., 2024). Computational CDS design must therefore navigate a large constrained sequence space while identifying sequences with desirable molecular properties.

Existing CDS-design methods broadly follow two paradigms. Explicit optimization methods, such as LinearDesign, search synonymous sequence space directly using predefined objectives (Zhang et al., 2023). Learned generative methods instead model natural coding sequences and use the learned distribution as a prior over plausible synonymous designs (Faizi et al., 2025; Zhang et al., 2024; Ren et al., 2024; Fallahpour et al., 2025). In this work, we explore discrete diffusion as a protein-conditioned generative formulation for CDS design. Its non-causal objective allows each codon to be predicted from the complete protein and bidirectional CDS context, while sequence tokens are generated in parallel through a small number of decoding steps.

A learned generative prior can then be adapted to application-specific objectives that may not be available as supervised training labels. Recent work has explored reinforcement learning for both CDS and full-length mRNA optimization, including approaches based on REINFORCE, actor–critic methods, and Group Relative Policy Optimization (GRPO) (Rajbanshi & Guruacharya, 2025; Du et al., 2026; Huang et al., 2025; Lin et al., 2026; Li et al., 2026; Shao et al., 2026).

Here, we study how a protein-conditioned discrete-diffusion prior can be adapted at two scales: first across diverse proteins to obtain a reusable reward-adapted policy, and then at test time for a fixed

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

protein target. We use normalized minimum free energy (MFE) as a non-differentiable sequence-level objective and experimentally evaluate selected designs using the NanoLuc reporter. Figure 1 summarizes the resulting three-stage framework: learning a protein-conditioned CDS prior, adapting it through global and test-time reinforcement learning, and experimentally evaluating the resulting designs. A detailed comparison with recent reward-guided RNA design methods and related diffusion-based adaptation approaches is given in Appendix A.

Our main contributions are:

1. **Prot2RNA, a protein-conditioned discrete diffusion model for CDS generation.** Prot2RNA outperforms existing methods in alignment with biologically meaningful sequence properties, generating profiles that closely mirror codon usage patterns in natural human CDSs and providing a strong foundation for protein-conditioned mRNA design.
2. **Reward guidance of a learned protein-conditioned CDS prior at two scales.** We show that Prot2RNA can be adapted toward a non-differentiable sequence-level objective through reinforcement finetuning across diverse proteins (**Prot2RNA-RFT**), and further shifted at test time for a fixed protein target using test-time reinforcement learning (**Prot2RNA-TTRL**), enabling optimization for properties like predicted CDS structural stability, that matter for therapeutic mRNA design but are inaccessible to standard supervised training.
3. **Experimental evaluation of reward-guided CDS designs.** We experimentally tested selected NanoLuc designs in HEK293T cells, showing that reward-guided adaptation yields measurably higher reporter output: our best design outperforms the experimentally validated GEMORNA and LinearDesign designs by 39% and 71%, respectively, demonstrating that reward-guided CDS optimization holds up in a real translational setting.

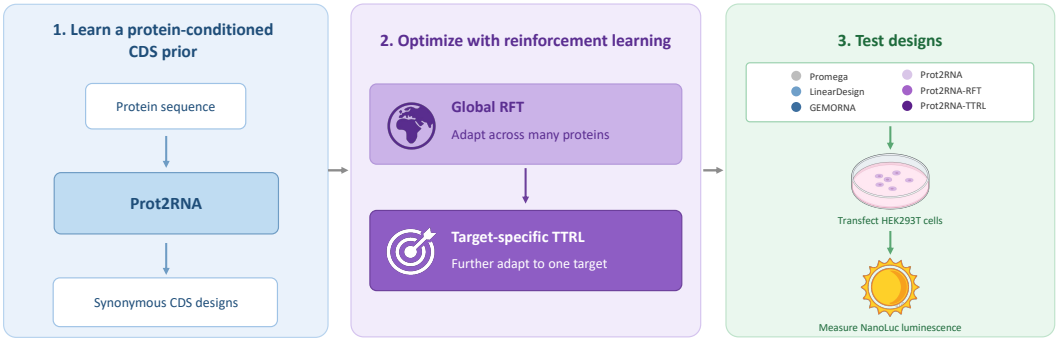

Figure 1: **Pipeline overview.** Prot2RNA learns a protein-conditioned prior over natural human CDSs (Stage 1). The model is then guided toward a sequence-level reward through global reinforcement finetuning (Prot2RNA-RFT) and further at test time for a fixed target (Prot2RNA-TTRL) (Stage 2). Selected designs are subsequently evaluated experimentally using NanoLuc (Stage 3).

2 METHOD

2.1 STAGE 1 (LEARN): PROT2RNA, A PROTEIN-CONDITIONED CDS PRIOR

Let $p = (p_1, \ldots, p_L)$ denote a protein sequence and $r = (r_1, \ldots, r_L)$ a corresponding CDS represented as codon tokens. Prot2RNA models the conditional distribution $p_\theta(r \mid p)$ over synonymous coding sequences using a Transformer encoder over a shared protein/codon vocabulary. Full architectural and training details are in Appendix B.

Pretraining and supervised finetuning. First, we pretrain the model on protein and CDS sequences independently using masked language modeling (MLM) with a fixed masking ratio of 15%,

$$\mathcal{L}_{\mathrm{MLM}}(\theta) = -\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \log p_\theta(x_i \mid x_{\setminus \mathcal{M}}), \tag{1}$$

where x is either a protein or a CDS sequence and $\mathcal{M}$ is the index set of masked tokens.

After pretraining, Prot2RNA is supervised finetuned on paired protein–CDS examples with the conditional discrete-diffusion objective of LLaDA (Nie et al., 2025): the protein stays fully visible, a masking ratio $t \sim U(0, 1)$ is drawn, codons are masked independently with probability t to give r_t, and the model reconstructs masked codons from the full protein and visible CDS context,

$$\mathcal{L}_{\mathrm{DLM}}(\theta) = -\mathbb{E}_{t,p,r}\left[\frac{1}{tL}\sum_{i\in\mathcal{M}} \log p_\theta(r_i \mid p, r_t)\right]. \tag{2}$$

This exposes the model to corruption levels ranging from nearly complete context to fully masked CDSs, matching the fully-masked conditional generation setting used at inference (Appendix Figure 6). We compare this DLM finetuning objective against matched NTP and MLM finetuning objectives under identical architecture, data, and training budget in Section 3.1.

Modality-aligned positional encoding. Standard rotary position embedding (RoPE, Su et al., 2024) assigns consecutive positional indices across the concatenated protein and CDS segments, despite the one-to-one correspondence between amino acid p_i and codon r_i. We instead assign corresponding protein and codon tokens the same positional index, $\phi(p_i) = \phi(r_i) = i$ for $i = 1, \ldots, L$, restarting CDS indices at the modality boundary after batch-max protein padding. This is a task-specific inductive bias for protein-conditioned CDS generation, and we quantify its effect against standard RoPE in Section 3.1.

Iterative sequence generation. Prot2RNA generates a CDS from a fully masked codon sequence in K iterative refinement steps. Following MaskGIT-style decoding (Chang et al., 2022), at each step, positions are ranked by confidence, defined as the maximum codon probability predicted by the model at each position, and a schedule-determined number of the highest-confidence positions is unmasked and fixed for subsequent steps, while the remaining positions stay masked and are reconsidered. This reduces the number of sequential decoding steps from L to $K \ll L$, which is advantageous when many complete rollouts are required for reinforcement finetuning. Unless otherwise stated, we use $K = 4$ with a cosine masking schedule (selected in Appendix C). For stochastic generation, codons are sampled using temperature scaling and nucleus sampling; deterministic evaluation uses greedy decoding.

2.2 Stage 2 (Optimize): global reinforcement finetuning and test-time reinforcement learning

We treat Prot2RNA as a policy $\pi_\theta(r \mid p)$ over CDSs given a protein p, and optimize it using sequence-level rewards evaluated after complete generation, which may be non-differentiable and need not be available at intermediate decoding steps. We use GRPO (Shao et al., 2024) to perform this optimization. For each protein p_b in a minibatch we sample a group of G candidates $\{r_{b,1}, \ldots, r_{b,G}\}$ and standardize rewards within the group,

$$\hat{A}_{b,i} = \frac{R_{b,i} - \mu_{R_b}}{\sigma_{R_b} + \delta}, \tag{3}$$

where μ_{R_b} and σ_{R_b} denote the mean and standard deviation of the G rewards sampled for protein p_b and $\delta = 10^{-8}$ is added for numerical stability, so that advantages compare synonymous candidates for the *same* protein.

For decoding iteration k, let s_k denote the current state (the conditioning protein and partially revealed CDS) and let $\mathcal{I}_k$ denote the codon positions revealed at that iteration. We denote the predicted distribution at position j by $\pi_\theta^{(j)}(\cdot \mid s_k)$ and the sampled codon at that position by $a_{k,j}$. The policy ratio is

$$\rho_\theta(a_{k,j} \mid s_k) = \frac{\pi_\theta^{(j)}(a_{k,j} \mid s_k)}{\pi_{\mathrm{old}}^{(j)}(a_{k,j} \mid s_k)}, \tag{4}$$

and the clipped GRPO objective is

$$\mathcal{L}_{\mathrm{GRPO}} = -\mathbb{E}_{b,i,k}\left[\frac{1}{|\mathcal{I}_k|}\sum_{j\in\mathcal{I}_k} \min\left(\rho_\theta(a_{k,j} \mid s_k)\hat{A}_{b,i}, \mathrm{clip}(\rho_\theta(a_{k,j} \mid s_k), 1-\epsilon_{\mathrm{low}}, 1+\epsilon_{\mathrm{high}})\,\hat{A}_{b,i}\right)\right]. \tag{5}$$

The same sequence-level advantage is assigned to all codon choices within a sampled CDS, while policy-ratio terms are computed for positions newly revealed at each decoding iteration. We regularize toward a frozen reference model π_{ref}, initialized from the pre-adaptation Prot2RNA policy, using

$$\widehat{D}_{\mathrm{KL}}(a_{k,j}, s_k) = \frac{\pi_{\mathrm{ref}}^{(j)}(a_{k,j} \mid s_k)}{\pi_\theta^{(j)}(a_{k,j} \mid s_k)} - \log \frac{\pi_{\mathrm{ref}}^{(j)}(a_{k,j} \mid s_k)}{\pi_\theta^{(j)}(a_{k,j} \mid s_k)} - 1, \tag{6}$$

and the final objective $\mathcal{L}_{\mathrm{RFT}} = \mathcal{L}_{\mathrm{GRPO}} + \beta\, \widehat{D}_{\mathrm{KL}}$, where smaller β permits larger departures from the reference policy toward the reward. Synonymous decoding is enforced throughout both global RFT and TTRL by masking logits for codons that do not encode the corresponding amino acid before sampling and renormalizing the resulting distribution over valid synonymous codons. Optimization can therefore change codon choice but not the encoded protein.

Terminal reward. We use normalized minimum free energy (MFE) as a non-differentiable sequence-level reward, $R(r) = -\frac{\mathrm{MFE}(r)}{L_{\mathrm{nt}}(r)}$, where $L_{\mathrm{nt}}(r)$ is the CDS length in nucleotides and MFE is computed with ViennaRNA (Lorenz et al., 2011). Normalization by sequence length enables comparison across CDSs of different lengths. Increased CDS secondary-structure stability has been associated with improved mRNA stability and sustained protein output in previous experimental studies (Mauger et al., 2019; Leppek et al., 2022; Bicknell et al., 2024).

We therefore use normalized MFE as a biologically motivated objective for evaluating reward-guided adaptation of Prot2RNA, while treating it as one design proxy rather than a direct predictor of protein expression.

Prot2RNA-RFT: global reinforcement finetuning. Prot2RNA-RFT performs reinforcement finetuning across a diverse set of 500 protein-cluster representatives sampled from the Prot2RNA training split (selected after clustering at 30% sequence identity to reduce redundancy), with 50 validation representatives selected analogously. This produces a single reusable policy, evaluated across held-out test proteins in Section 3.2. Full optimization details are in Appendix D.

Prot2RNA-TTRL: Test-Time Reinforcement Learning for a fixed target. Global RFT yields a reusable policy across proteins. Prot2RNA-TTRL further adapts this policy at test time for a fixed protein target using Test-Time Reinforcement Learning (TTRL; Zuo et al., 2026). Starting from the highest-reward global RFT checkpoint, we continue GRPO updates using only the target protein and the same normalized-MFE reward, retaining the highest-reward sequence encountered during optimization as the final candidate. We apply Prot2RNA-TTRL to NanoLuc in this work. Full staged hyperparameter search and selection details are provided in Appendix E.

2.3 Data and evaluation

Dataset and splits. We train on 118,224 human protein–CDS pairs derived from NCBI RefSeq. Train, validation, and test sets are constructed at the protein-cluster level using MMseqs2 (Steinegger & Söding, 2017; 2018) to reduce homology leakage across splits; the held-out test set contains 3,004 pairs. Proteins similar to prospective experimental targets, including NanoLuc, are excluded from training. Full preprocessing and split construction are in Appendix F.

Baselines and metrics. We compare Prot2RNA against CAI-max, a codon-optimization baseline that selects the most frequent human synonymous codon per amino acid; LinearDesign (Zhang et al., 2023); Generative CodonBERT (Ren et al., 2024); CodonTransformer (Fallahpour et al., 2025); Trias (Faizi et al., 2025); and GEMORNA (Zhang et al., 2024). Recent RL-based RNA-design methods (Rajbanshi & Guruacharya, 2025; Du et al., 2026; Huang et al., 2025; Lin et al., 2026; Li et al., 2026; Shao et al., 2026) are not included as direct quantitative baselines. Several do not provide executable implementations, and others optimize a separate policy for each target protein rather than learning a reusable policy across proteins, making a controlled comparison to Prot2RNA-RFT difficult. We therefore evaluate the effect of reinforcement finetuning relative to the same pretrained Prot2RNA policy, and discuss these methods in Appendix A. We evaluate generated sequences using codon accuracy, CAI, nucleotide composition, and normalized MFE, and assess

population-level similarity to natural CDSs using Fréchet Transcript Distance (FTD) in a transcript-language-model embedding space. Full metric definitions and FTD methodology, including finite-sample calibration, are provided in Appendices G and H.

3 RESULTS

3.1 STAGE 1: A DISCRETE-DIFFUSION PRIOR WITH A PROTEIN-CODON ALIGNED POSITIONAL ENCODING

Under matched architecture, data, and compute, diffusion finetuning gives the strongest conditional CDS model among the training objectives we evaluate on the validation set (Table 1): MLM→DLM reaches mean codon accuracy 0.688, compared with 0.554 for NTP→NTP and 0.526 for MLM→MLM (full variant definitions in Appendix I). FTD provides a complementary population-level measure of similarity to natural CDSs. NTP→NTP and both DLM variants achieve substantially lower FTD than MLM→MLM. However, their FTD values are close to the finite-sample baseline of approximately 3 at this validation-set size ($n \approx 1{,}700$). We therefore use codon accuracy as the primary metric for selecting the conditional generation objective.

Table 1: Effect of pretraining and finetuning objectives on codon accuracy and FTD in the validation set.

Objective	Accuracy ↑	FTD ↓
NTP → NTP	0.554 ± 0.172	5.21
MLM → MLM	0.526 ± 0.073	18.97
DLM → DLM	0.664 ± 0.190	**3.86**
MLM → DLM	$\mathbf{0.688 \pm 0.203}$	4.62

Table 2: Effect of modality-aligned positional encoding and middle padding on codon accuracy in the validation set.

Configuration	Accuracy ↑
Aligned + padding	**0.688**
Standard + padding	0.521
Standard, no padding	0.510

The modality-aligned positional encoding is at least as important as the training objective: with the training strategy and decoding procedure fixed, replacing it with standard RoPE reduces mean codon accuracy from 0.688 to 0.521, and removing the intermediate protein-padding as well reduces it only slightly further, to 0.510 (Table 2). The much larger drop from removing positional alignment than from removing padding indicates that explicit protein–codon positional correspondence, rather than the padding layout, accounts for most of this gain (full configuration details in Appendix J).

On the held-out test set, Prot2RNA reaches mean codon accuracy 0.647 ± 0.186, compared with 0.559 ± 0.131 for the strongest external baseline by this metric, Trias (Table 3). As supporting evidence that generated sequences remain close to natural human CDS distributions, Prot2RNA reaches an FTD of 3.69, lower than all evaluated baselines against the combined validation and test population ($n \approx 4{,}700$; estimated finite-sample baseline ~ 1.1; full calibration in Appendix H), and shows smaller shifts than several baselines across CAI, GC content, and uridine fraction (Figure 2). Prot2RNA is evaluated here without synonymous-codon constraints during decoding, yet all generated test-set sequences translate exactly to the target protein, indicating that protein preservation emerges from the conditional objective itself rather than from decoding constraints (per-baseline mismatch rates under the same unconstrained setting are reported in Appendix K).

Table 3: **Comparison with CDS-design baselines.** Codon accuracy is evaluated on the held-out test set; FTD is computed against the combined validation and test population ($n \approx 4{,}700$).

Method	Codon accuracy (mean ± std)	Median	FTD ↓
LinearDesign ($\lambda = 0$)	0.398 ± 0.063	0.398	13.68
GEMORNA	0.451 ± 0.091	0.441	18.65
LinearDesign ($\lambda = 3$)	0.446 ± 0.089	0.451	10.67
CodonBERT	0.476 ± 0.119	0.479	34.03
CAI-max	0.458 ± 0.121	0.464	73.54
CodonTransformer	0.502 ± 0.086	0.505	11.64
Trias	0.559 ± 0.131	0.552	17.00
Prot2RNA	$\mathbf{0.647 \pm 0.186}$	**0.600**	**3.69**

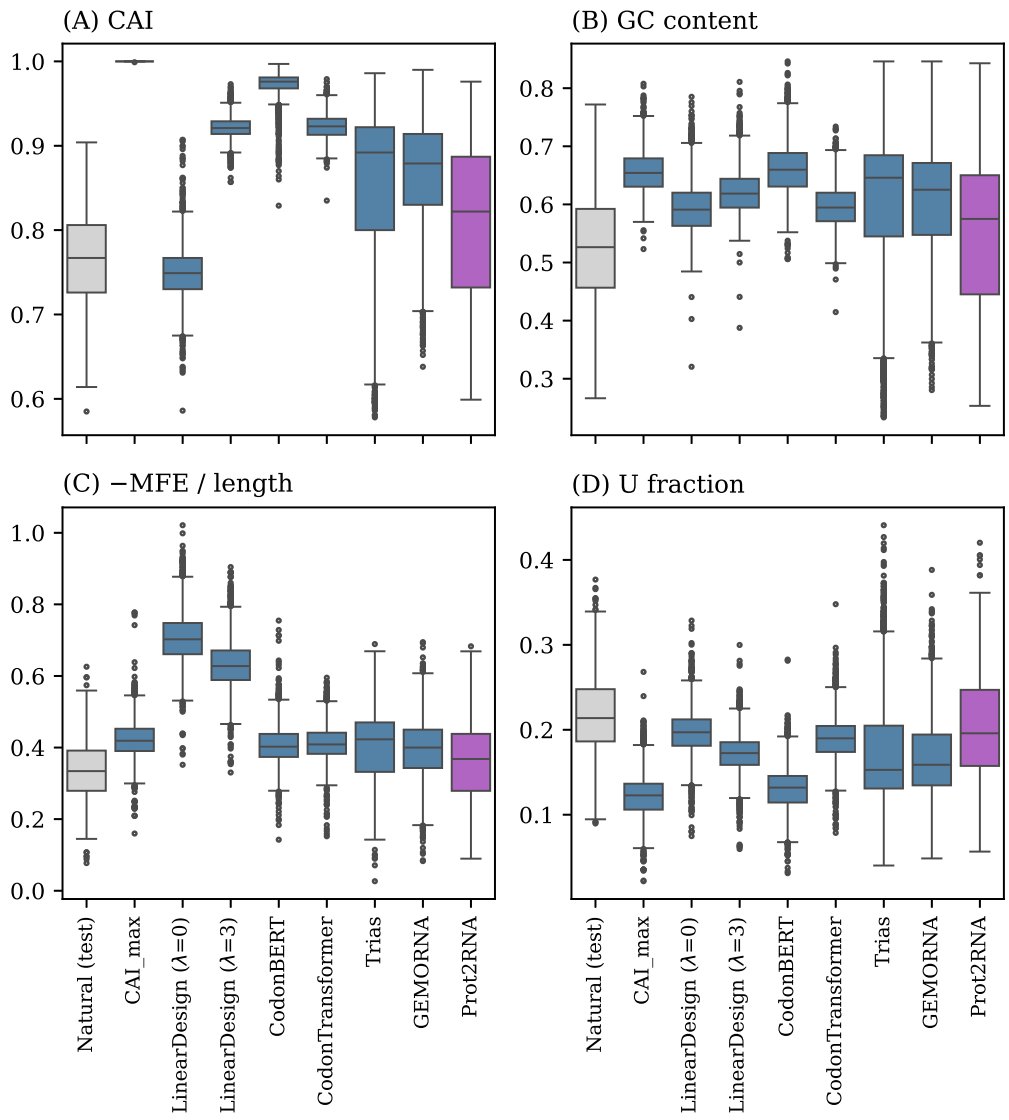

Figure 2: Biological properties of generated sequences for Prot2RNA and baselines, compared to natural human CDS (test set). (A) CAI; (B) GC content; (C) −MFE/length; (D) uridine fraction.

Together, these results establish Prot2RNA as a strong protein-conditioned CDS prior. Discrete-diffusion finetuning improves conditional generation over matched NTP and MLM objectives, while explicitly aligning protein and codon positions provides a substantial additional gain. The resulting model generates CDSs that more closely resemble natural human coding sequences than the evaluated baselines while preserving the encoded protein without requiring synonymous constraints during decoding. Thus, Prot2RNA provides a foundation for further optimizing CDSs for specific applications.

3.2 Stage 2.1: Global reinforcement finetuning shifts normalized MFE across held-out proteins

Prot2RNA-RFT consistently shifts normalized MFE in the intended direction across held-out test proteins: median normalized MFE increases from 0.355 for Prot2RNA to $0.465 - 0.479$ across the evaluated KL coefficients β (Figure 3A). Weaker KL regularization permits stronger reward optimization but reduces within-protein sequence diversity (Figure 3B). This shift is not reproduced by sampling more candidates from the pre-RFT model: best-of-32 selection from Prot2RNA reaches

mean normalized MFE 0.354, compared with $0.468-0.482$ under ordinary stochastic sampling from the RFT policies (Figure 4A). This demonstrates that RFT shifts the sampled distribution toward higher-reward sequences, rather than merely increasing the probability of finding rare high-reward samples already accessible from the pretrained Prot2RNA policy (full best-of-N methodology in Appendix N). This shift in the learned conditional distribution is further reflected in the FTD values, with sequences optimized for predicted structural stability moving away from the natural-CDS distribution. Larger FTD shifts are generally observed under weaker KL regularization (Appendix L).

The reward shift is observed consistently across independent training runs: three independently trained $\beta = 0.005$ policies (seeds 42/43/44) produce closely matching mean normalized MFE and FTD (Table 4). Coordinated secondary shifts in GC content, uridine fraction, and CAI accompany this optimization but are not uniformly monotonic in β, since only normalized MFE is directly optimized.

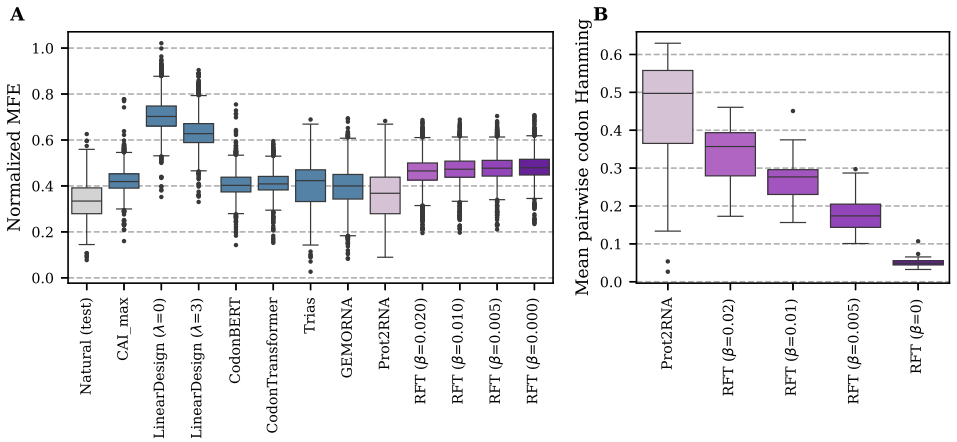

Figure 3: Effect of Prot2RNA-RFT on normalized MFE and sequence diversity. (A) Normalized MFE for natural sequences, baseline methods, Prot2RNA, and RFT policies on test-set proteins. (B) Within-protein sequence diversity under stochastic decoding across KL coefficients for 30 validation proteins.

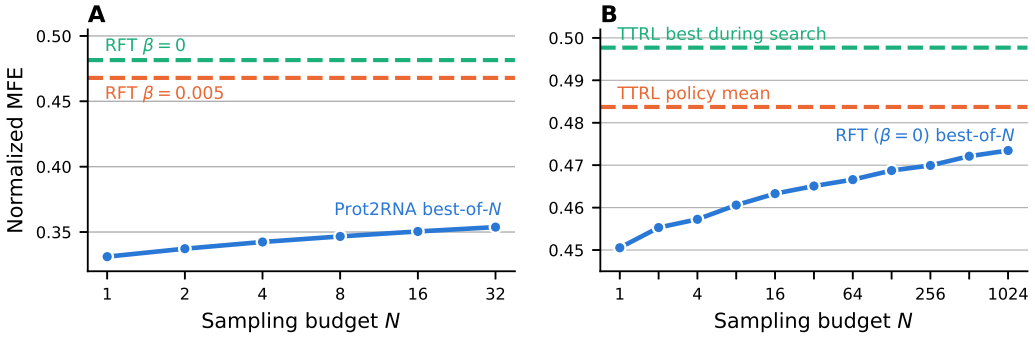

Figure 4: **Reward adaptation shifts the policy beyond increased sampling from the preceding policy.** (A) Best-of-N sampling from Prot2RNA compared with mean reward under global RFT policies. (B) Best-of-N sampling from the global RFT policy compared with the mean reward under the final NanoLuc TTRL policy and the highest-reward sequence identified during TTRL.

3.3 Stage 2.2: Test-time reinforcement learning for NanoLuc

For NanoLuc, Prot2RNA-TTRL is initialized from the global RFT checkpoint trained without KL regularization ($\beta = 0$), which achieved the highest normalized-MFE reward among the global RFT

Table 4: Prot2RNA-RFT ($\beta = 0.005$) across three random seeds (test set).

Seed	Normalized MFE ↑	FTD ↓
42	0.4796 ± 0.0566	57.45
43	0.4795 ± 0.0567	56.51
44	0.4796 ± 0.0570	58.25

variants. We then perform a staged hyperparameter search over learning rate, number of unfrozen Transformer blocks, group size G, sampling temperature, and decoding configuration (Appendix E).

To distinguish improvement through policy adaptation from increased sampling of the starting RFT policy, we compare TTRL against best-of-N selection from $1,024$ candidates generated by the initial RFT policy (Figure 4B; full methodology in Appendix N.2). Best-of-N reward increases with the sampling budget but reaches only approximately 0.474 at $N = 1{,}024$. In contrast, ordinary stochastic sampling from the final TTRL policy achieves a mean normalized MFE of approximately 0.484 across 64 samples, exceeding even the best reward found among $1,024$ samples from its starting RFT policy. The highest-reward sequence identified during the complete TTRL search reaches 0.4977 and is evaluated experimentally in Section 3.4. These results show that target-specific policy adaptation provides additional reward improvement beyond substantially increasing the sampling budget of the global RFT policy.

Together, global RFT and NanoLuc TTRL demonstrate complementary levels of reward adaptation: RFT shifts a reusable protein-conditioned policy across held-out proteins, while TTRL further specializes that policy once a particular target is known.

3.4 Stage 3: Experimental evaluation of NanoLuc coding-sequence designs

We evaluated selected synonymous CDS designs experimentally using NanoLuc, a compact luciferase reporter whose protein output is quantified by luminescence. Constructs were tested in HEK293T cells 72 h after transfection, with two replicate wells per construct, and included the Promega reference, two negative controls (a deliberately low-GC% control and Yellowstone, a published low-output NanoLuc CDS from Leppek et al., 2022), LinearDesign (Zhang et al., 2023), GEMORNA (Zhang et al., 2024), Prot2RNA, two Prot2RNA-RFT designs ($\beta = 0.005$ and $\beta = 0$), and Prot2RNA-TTRL (Figure 5). All constructs encode the same NanoLuc protein, and the Prot2RNA/RFT/TTRL candidates were selected before assay results were measured. Full construct-selection and experimental details are in Appendix O.

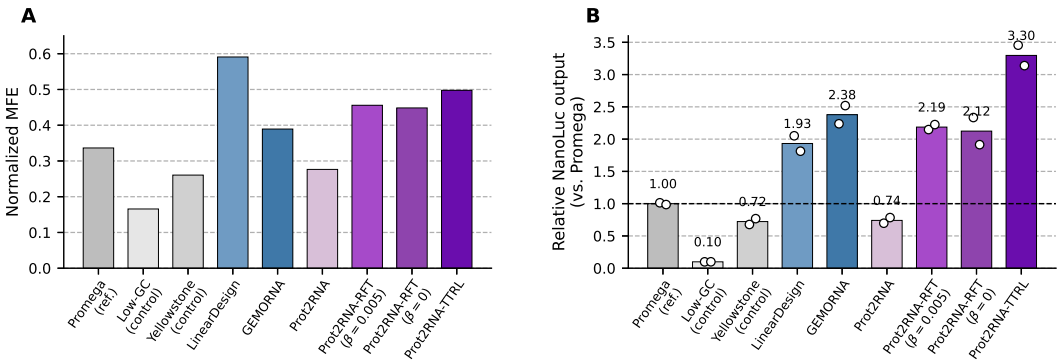

Figure 5: **Experimental evaluation of synonymous NanoLuc CDS designs.** (A) Normalized MFE of the tested constructs. (B) NanoLuc output measured in HEK293T cells 72 h after transfection, normalized to the Promega reference. Bars show the mean of two replicate wells, with individual replicate measurements shown as circles.

Synonymous CDS choice produced substantial differences in measured reporter output despite all constructs encoding the same protein. The TTRL-selected design gave the highest measured output in the tested panel, reaching $3.30\times$ the Promega reference, 39% higher than GEMORNA ($2.38\times$)

and 71% higher than LinearDesign ($1.93\times$). The two global RFT designs reached 2.12–$2.19\times$, whereas the pre-RFT Prot2RNA design reached $0.74\times$. The low-GC% and Yellowstone negative controls reached $0.10\times$ and $0.72\times$, respectively.

Within the tested Prot2RNA lineage, reward adaptation was accompanied by increasing reporter output: normalized MFE increased from 0.276 for Prot2RNA to 0.448–0.456 after global RFT and 0.498 after TTRL, while reporter output increased from $0.74\times$ to 2.12–$2.19\times$ and finally $3.30\times$ the Promega reference. Thus, both stages of reward adaptation produced designs with higher measured reporter output than the pre-RFT Prot2RNA design in this experiment.

Importantly, this relationship did not extend monotonically across the full panel. LinearDesign directly incorporates folding stability into its design objective, and the tested LinearDesign construct had the highest normalized MFE among all tested constructs (0.591), yet produced lower reporter output than GEMORNA and the RFT/TTRL designs. Normalized MFE therefore captures only one component of CDS design rather than directly predicting reporter output. Other properties captured by learned sequence models, including natural codon-usage patterns and broader CDS statistics, may also contribute to reporter output. In addition, increased CDS structure may stabilize the transcript while sufficiently stable secondary structures can impede translation and require unwinding during elongation (Waudby et al., 2019; Bao & Ermolenko, 2021). Further experiments would be required to determine which sequence properties and mechanisms explain the observed differences between the tested constructs.

4 Discussion

Prot2RNA proposes a framework for CDS design in which learning and optimization are separated but complementary. Rather than optimizing each protein independently from scratch, Prot2RNA first learns a reusable protein-conditioned distribution over natural human CDSs. This prior can then be shifted globally toward a sequence-level objective through reinforcement finetuning and further specialized through test-time reinforcement learning once a particular protein target is known. Global RFT consistently shifts normalized MFE across held-out proteins, while for NanoLuc, TTRL provides additional reward improvement beyond substantially increasing the sampling budget of the global RFT policy. The resulting TTRL-selected NanoLuc design produced the highest measured reporter output in the tested panel, reaching $3.30\times$ the Promega reference. Together, these results establish a two-scale strategy for reward-guided CDS design: learn a general protein-conditioned sequence prior once, adapt it globally toward a desired property, and further specialize it when a specific target warrants additional optimization.

The NanoLuc experiment also highlights why reward choice is central to this framework. Within the Prot2RNA lineage, global and target-specific reward adaptation increased normalized MFE and were accompanied by higher measured reporter output. Across the full experimental panel, however, the construct with the highest normalized MFE did not produce the highest reporter output. This shows that optimizing predicted structural stability alone does not provide a complete objective for protein output and backs the interplay between different sequence properties captured by the generative priors as an important piece of the whole puzzle. This distinction creates a natural direction for extending Prot2RNA: the same reward-adaptation framework can in principle incorporate objectives that capture complementary aspects of CDS function, including translation efficiency and transcript stability, or combine multiple properties into a single design objective. Experimentally informed rewards could further reduce the gap between computational optimization and the biological phenotype ultimately of interest.

The experimental evidence in this study is limited to a single NanoLuc experiment with two replicate wells per construct, and TTRL was evaluated for NanoLuc only. Additional protein targets and independent experimental replicates are therefore needed to determine how broadly the observed target-specific and experimental results generalize. More broadly, evaluating alternative and multi-objective rewards across diverse targets will be important for determining which computational objectives transfer most reliably to biological outcomes. Prot2RNA provides a platform for this setting in which the protein-conditioned prior is retained while the optimization objective can be changed according to the requirements of the design task.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experimental design and to assist with interpretation of results. We did not use generative AI tools to generate synthetic datasets, formulate mathematical claims or proofs, or perform qualitative or thematic data analysis; the remaining required-disclosure tasks were not applicable to this work.

Additionally, we used generative AI tools for software development and debugging, literature exploration, and drafting and refining manuscript text. All AI-assisted work was reviewed by the authors. AI-assisted code was reviewed and tested for correctness, and literature and scientific claims were checked against the original sources. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available biological sequence data and in vitro experiments with established cell lines. It does not involve human participants, identifiable personal data, or animal experiments.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of dataset construction and splitting, model architecture and training, iterative decoding, reinforcement finetuning, and test-time reinforcement learning in the appendices. The appendices additionally specify evaluation metrics, baseline configurations, best-of-N sampling procedures, random-seed analyses, and the experimental protocol used for NanoLuc validation. Training and evaluation hyperparameters are reported throughout the corresponding sections to facilitate reproduction of the computational experiments. Anonymous source code, model weights, and the dataset used in this work are available at `https://figshare.com/s/9132042e385a7324ca64`.

REFERENCES

Sarah Alamdari, Nitya Thakkar, Rianne Van Den Berg, Neil Tenenholtz, Robert Strome, Alan M Moses, Alex X Lu, Nicolò Fusi, Ava P Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. *BioRxiv*, pp. 2023–09, 2023.

Oliwia Andries, Séan Mc Cafferty, Stefaan C De Smedt, Ron Weiss, Niek N Sanders, and Tasuku Kitada. N1-methylpseudouridine-incorporated mRNA outperforms pseudouridine-incorporated mRNA by providing enhanced protein expression and reduced immunogenicity in mammalian cell lines and mice. *Journal of Controlled Release*, 217:337–344, 2015.

Chen Bao and Dmitri N Ermolenko. Ribosome as a translocase and helicase. *Biochemistry (Moscow)*, 86(8):992–1002, 2021.

Alicia A Bicknell, David W Reid, Marissa C Licata, Adriana K Jones, Yi Min Cheng, Mengying Li, Chiaowen Joyce Hsiao, Christopher S Pepin, Mihir Metkar, Yevgen Levdansky, et al. Attenuating ribosome load improves protein output from mRNA by limiting translation-dependent mRNA decay. *Cell reports*, 43(4), 2024.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. MaskGIT: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Shiyi Du, Gün Kaynar, Jiayi Li, Zhaoyi You, Shijie Tang, and Carl Kingsford. CodonRL: Multi-Objective Codon Sequence Optimization Using Demonstration-Guided Reinforcement Learning. *bioRxiv*, 2026.

Marjan Faizi, Helen Sakharova, and Liana F Lareau. A generative language model decodes contextual constraints on codon choice for mRNA design. *bioRxiv*, pp. 2025–05, 2025.

Adibvafa Fallahpour, Vincent Gureghian, Guillaume J Filion, Ariel B Lindner, and Amir Pandi. CodonTransformer: a multispecies codon optimizer using context-aware neural networks. *Nature Communications*, 16(1):3205, 2025.

Yijia Fan, Jing Yang, Mingyu Liu, Kaitong Cai, Jian Wang, Keze Wang, and Jusheng Zhang. Reinforcement learning for diffusion llms via energy-based gibbs alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 45938–45948, 2026.

Fabian Hia, Sheng Fan Yang, Yuichi Shichino, Masanori Yoshinaga, Yasuhiro Murakawa, Alexis Vandenbon, Akira Fukao, Toshinobu Fujiwara, Markus Landthaler, Tohru Natsume, et al. Codon bias confers stability to human mRNAs. *EMBO reports*, 20(11):e48220, 2019.

Jiaqi Huang, Ning Feng, Huarong Bai, Yi Fang, Xiaojian Liu, Shengfan Wang, Junchi Yan, Hong-Bin Shen, Zilong Qiu, Ye Yuan, et al. RNAJog: Fast Multi-objective RNA Optimization with Autoregressive Reinforcement Learning. *bioRxiv*, pp. 2025–08, 2025.

Katalin Karikó, Hiromi Muramatsu, Frank A Welsh, János Ludwig, Hiroki Kato, Shizuo Akira, and Drew Weissman. Incorporation of pseudouridine into mRNA yields superior nonimmunogenic vector with increased translational capacity and biological stability. *Molecular therapy*, 16(11): 1833–1840, 2008.

Grzegorz Kudla, Leszek Lipinski, Fanny Caffin, Aleksandra Helwak, and Maciej Zylicz. High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS biology*, 4(6): e180, 2006.

Kathrin Leppek, Gun Woo Byeon, Wipapat Kladwang, Hannah K Wayment-Steele, Craig H Kerr, Adele F Xu, Do Soon Kim, Ved V Topkar, Christian Choe, Daphna Rothschild, et al. Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. *Nature communications*, 13(1):1536, 2022.

Sizhen Li, Saeed Moayedpour, Ruijiang Li, Michael Bailey, Saleh Riahi, Lorenzo Kogler-Anele, Milad Miladi, Jacob Miner, Fabien Pertuy, Dinghai Zheng, et al. CodonBERT large language model for mRNA vaccines. *Genome research*, 34(7):1027–1035, 2024.

Sizhen Li, Paul Chauvin, Ofek Gross, Michael Bailey, and Sven Jager. mRNA-GPT: A Generative Model for Full-Length mRNA Design and Optimization. *bioRxiv*, pp. 2026–03, 2026.

Shenggeng Lin, Hong Tan, Keyao Wang, Ruixuan Wang, Hongxia Wang, Tong Zhu, and Yi Xiong. RNARL: reinforcement learning-driven unified generative framework for multi-objective RNA codon design. *Genome Biology*, 2026.

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*, 6:1–14, 2011.

Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam, 2017.

David M Mauger, B Joseph Cabral, Vladimir Presnyak, Stephen V Su, David W Reid, Brooke Goodman, Kristian Link, Nikhil Khatwani, John Reynders, Melissa J Moore, et al. mRNA structure regulates protein expression through changes in functional half-life. *Proceedings of the National Academy of Sciences*, 116(48):24075–24083, 2019.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models, 2025.

Norbert Pardi, Michael J Hogan, Frederick W Porter, and Drew Weissman. mRNA vaccines—a new era in vaccinology. *Nature reviews Drug discovery*, 17(4):261–279, 2018.

Sawan Patel, Sophia Tang, Yinuo Zhang, Pranam Chatterjee, and Sherwood Yao. Multi-objective-guided generative design of mRNA with therapeutic properties. In *ICML 2025 Workshop on Scaling Up Intervention Models*, 2025.

Rafael Josip Penić, Tin Vlašić, Roland G Huber, Yue Wan, and Mile Šikić. RiNALMo: general-purpose RNA language models can generalize well on structure prediction tasks. *Nature Communications*, 16(1):5671, 2025.

Binita Rajbanshi and Anuj Guruacharya. codonGPT: reinforcement learning on a generative language model enables scalable mRNA design. *Nucleic Acids Research*, 53(22):gkaf1345, 2025.

Zilin Ren, Lili Jiang, Yaxin Di, Dufei Zhang, Jianli Gong, Jianting Gong, Qiwei Jiang, Zhiguo Fu, Pingping Sun, Bo Zhou, et al. CodonBERT: a BERT-based architecture tailored for codon optimization using the cross-attention mechanism. *Bioinformatics*, 40(7):btae330, 2024.

Peter Rice, Ian Longden, and Alan Bleasby. Emboss: the european molecular biology open software suite. *Trends in genetics*, 16(6):276–277, 2000.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Ugur Sahin, Katalin Karikó, and Özlem Türeci. mRNA-based therapeutics—developing a new class of drugs. *Nature reviews Drug discovery*, 13(10):759–780, 2014.

Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Neural Information Processing Systems (NeurIPS)*, pp. 130136–130184, 2024.

Anirban Sarkar, Alejandra Duran, Yiyang Yu, Da-Wei Lin, Yijie Kang, Nirali Somia, Pablo Mantilla, Jessica Zhou, Masayuki Nagai, Ziqi Tang, et al. Designing DNA with tunable regulatory activity using discrete diffusion. *bioRxiv*, pp. 2024–05, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Zixi Shao, Tao Wang, Yibei Xiao, and Tianyi Huang. Protein-Conditioned Multi-Objective Reinforcement Learning for Full-Length mRNA Design. *arXiv preprint arXiv:2605.01513*, 2026.

P.M. Sharp and W.H. Li. The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, 15(3):1281–1295, 1987. doi: 10.1093/nar/15.3.1281.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, November 2017.

Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nature Communications*, 9(1):2542, June 2018.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

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*, 2024a.

Xinyou Wang, Zaixiang Zheng, Fei YE, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. In *Forty-first International Conference on Machine Learning*, 2024b.

Christopher A Waudby, Christopher M Dobson, and John Christodoulou. Nature and regulation of protein folding on the ribosome. *Trends in biochemical sciences*, 44(11):914–926, 2019.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

He Zhang, Liang Zhang, Ang Lin, Congcong Xu, Ziyu Li, Kaibo Liu, Boxiang Liu, Xiaopin Ma, Fanfan Zhao, Huiling Jiang, et al. Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature*, 621(7978):396–403, 2023.

He Zhang, Hailong Liu, Yushan Xu, Yiming Liu, Jia Wang, Yan Qin, Haiyan Wang, Lili Ma, Zhiyuan Xun, Timothy K Lu, et al. Deep generative models generate mRNA sequences with enhanced translation capacity and stability. *bioRxiv*, pp. 2024–06, 2024.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Xinwei Long, Ermo Hua, Biqing Qi, et al. TTRL: Test-time reinforcement learning. *Advances in Neural Information Processing Systems*, 38:131459–131483, 2026.

A Related Work

Protein-conditioned CDS design. Existing CDS-design methods span explicit optimization and learned conditional generation. LinearDesign (Zhang et al., 2023) directly searches synonymous sequence space to balance RNA folding stability and codon usage. Learned approaches instead train on natural coding sequences and differ in how they adapt the learned distribution to a target host or desired sequence profile. Generative CodonBERT (Ren et al., 2024) and CodonTransformer (Fallahpour et al., 2025) use masked modeling objectives, while Trias (Faizi et al., 2025) and GEMORNA (Zhang et al., 2024) use autoregressive generation. Existing learned CDS generators therefore rely primarily on masked reconstruction or autoregressive factorization, motivating discrete diffusion as an alternative and better-aligned formulation for protein-conditioned synonymous sequence generation.

Discrete diffusion for biological sequence generation. Discrete diffusion provides a non-causal alternative to autoregressive sequence modeling, generating sequences through iterative denoising while conditioning on bidirectional context (Nie et al., 2025; Sahoo et al., 2024). It has achieved strong empirical results in protein design, including protein sequence generation with EvoDiff and DPLM (Alamdari et al., 2023; Wang et al., 2024b), and regulatory DNA design (Sarkar et al., 2026). Most relevant to us, mRNAutilus (Patel et al., 2025) uses masked discrete diffusion with inference-time search for multi-objective mRNA design: a pretrained model's weights are guided at sampling time rather than adapted through reinforcement learning, and no wet-lab validation is reported. Our work differs in adapting the diffusion policy's parameters directly via reinforcement learning at two scales, global and target-specific, and in grounding the result with a wet-lab reporter assay.

Discrete diffusion and reward alignment. Reward-guided adaptation has also been studied for diffusion language models more broadly (Fan et al., 2026). Most directly relevant, DRAKES (Wang et al., 2024a) fine-tunes pretrained discrete diffusion models toward downstream rewards on DNA enhancer activity and protein stability (inverse folding), by making the diffusion trajectory differentiable via a Gumbel-softmax relaxation and backpropagating the reward through it. Our reward, normalized MFE from an external folding routine, is not made differentiable; we instead optimize it with GRPO over discrete rollouts under a synonymous-codon constraint that structures the action space more tightly than free-alphabet DNA or protein generation, and we introduce a second, target-specific adaptation stage (TTRL) on top of global reward finetuning, which DRAKES does not consider.

Reward-guided RNA design. A substantial body of recent work has applied reinforcement learning to CDS and full-length mRNA optimization. codonGPT (Rajbanshi & Guruacharya, 2025) and CodonRL (Du et al., 2026) optimize separate policies for individual protein targets using REINFORCE and demonstration-guided deep Q-learning, respectively; CodonRL reports quantitative gains over GEMORNA on CAI, MFE, and uridine content across a benchmark of human proteins. RNAJog (Huang et al., 2025) uses autoregressive actor–critic optimization of MFE together with controllable codon-usage and GC objectives, and reports its own wet-lab validation in a mouse influenza HA mRNA vaccine model. RNARL (Lin et al., 2026) uses GRPO for protein-conditioned CDS generation with a combined MFE–CAI reward; mRNA-GPT (Li et al., 2026) and ProMORNA (Shao et al., 2026) extend reward-guided optimization to untranslated regions or full-length mRNA design using learned predictors of stability and translation efficiency, evaluated in silico by the authors' own account. Relative to this body of work, our contribution is (a) a strong discrete-diffusion foundation for protein-conditioned CDS design, (b) the discrete-diffusion policy adapted at two distinct scales, global and target-specific, and (c) a wet-lab comparison against two extensively-validated baselines (LinearDesign and GEMORNA).

B MODEL ARCHITECTURE AND TRAINING HYPERPARAMETERS

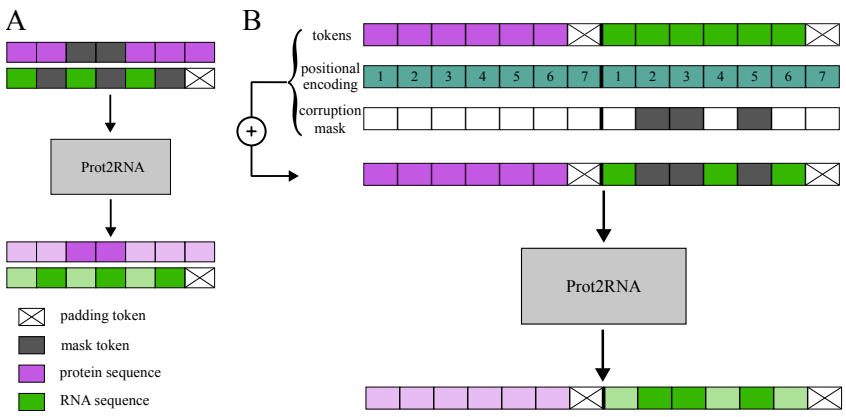

Figure 6: **Prot2RNA training.** (A) MLM pretraining models protein and CDS sequences independently. (B) Conditional diffusion finetuning keeps the protein fully visible while masking only the CDS. Protein sequences are padded to the batch-maximum length before the CDS is appended, and corresponding amino-acid and codon positions share the same positional index.

Alphabet and tokenization. The model employs a shared vocabulary of 21 amino acid tokens and 64 RNA codon tokens. Modality-specific special tokens denote sequence boundaries: `[BOS_P]` and `[EOS_P]` for proteins, `[BOS_R]` and `[EOS_R]` for RNAs. Common tokens, such as `[MASK]` and `[PAD]`, are shared across both modalities. Protein sequences are always terminated with the "$\star$" symbol, corresponding to the stop codon in RNA, to preserve amino-acid/codon alignment.

Architecture. Prot2RNA uses the RiNALMo (Penić et al., 2025) Transformer encoder backbone in its Micro configuration: 12 Transformer blocks, embedding dimension 480, 20 attention heads, and approximately 33.5 million trainable parameters, with LayerNorm replaced by RMSNorm (Zhang & Sennrich, 2019) for efficiency.

Two-stage training. Both stages use AdamW (Loshchilov & Hutter, 2017) with weight decay 0.01 and linear warmup followed by cosine annealing, for 50,000 optimization steps each on $4\times$NVIDIA A100 (40 GB) GPUs. During MLM pretraining, protein and coding sequences are treated as independent training examples with a fixed 15% of tokens masked at random; the learning rate is warmed up to 1×10^{-4} over the first 2,000 steps, then annealed to 1×10^{-5}; effective batch size is 1,536. During DLM finetuning, protein and coding sequences are concatenated into a single input with masking applied only to the coding sequence, using masking ratio $t \sim \mathcal{U}(0, 1)$; the learning rate is warmed up to 3×10^{-5} over the first 500 steps, then annealed to 8×10^{-6}; effective batch size is 512.

Sequence length constraints. During pretraining, protein and CDS sequences are processed independently with a context length of 1,024 tokens; longer sequences are randomly cropped to 1,022 tokens per epoch, following ESM-1b (Rives et al., 2021). During conditional finetuning, protein and CDS sequences are concatenated into a single input with a maximum context length of 2,048 tokens, so training is restricted to pairs with protein/CDS length below 1,022 amino-acid/codon positions to preserve the one-to-one protein–codon alignment.

C ITERATIVE DECODING AND MASKING SCHEDULES

Generation procedure. Generation begins from a fully masked coding sequence $r^{(0)} = (\texttt{[MASK]}, \ldots, \texttt{[MASK]})$, with length determined by the target protein. At each decoding iteration $k = 1, \ldots, K$: (1) the partially reconstructed sequence $r^{(k-1)}$ is concatenated with the protein

sequence p and passed through Prot2RNA, which predicts a distribution over codons at every currently masked position; (2) codons are sampled, or greedily selected for deterministic decoding, at every masked position; (3) only a schedule-determined subset of these newly predicted positions is retained, with the remainder re-masked and reconsidered in later iterations, according to a masking schedule $\gamma(k/K)$. This follows MaskGIT-style iterative decoding (Chang et al., 2022).

Effect of the number of decoding iterations and masking schedule. We evaluated five masking schedules (cubic, square, cosine, linear, square-root) across different numbers of decoding iterations on the validation set, measuring codon accuracy, normalized MFE, and FTD under deterministic decoding (Figure 7). The number of decoding iterations has a clearer effect than the masking schedule: codon accuracy decreases modestly as more iterations are added, while normalized MFE and FTD both increase, with most of the change occurring within the first few iterations. Notably, the increase in normalized MFE emerges without MFE being used during Prot2RNA training, suggesting that iterative refinement progressively favors sequence features associated with greater predicted structural stability. Differences between masking schedules are comparatively small, and no schedule consistently performs best across all metrics. We therefore use $K = 4$ with the cosine schedule as the default configuration for Prot2RNA and global RFT; TTRL additionally explores these decoding hyperparameters during test-time adaptation (Appendix E).

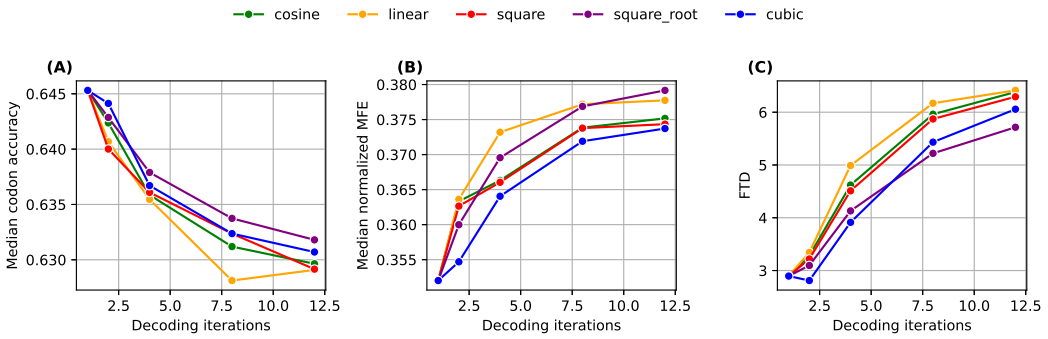

Figure 7: Effect of the number of decoding iterations and masking schedule on Prot2RNA. (A) Median codon accuracy, (B) median normalized MFE, and (C) FTD on the validation set.

D Reinforcement finetuning: objective and training details

Reinforcement finetuning initializes from the finetuned Prot2RNA checkpoint and runs for 1,000 update steps using AdamW with a constant learning rate of 1×10^{-6}, weight decay 0.1, and $(\beta_1, \beta_2) = (0.9, 0.95)$. Training uses bfloat16 mixed precision and seed 42 (see Appendix M for seed robustness). We use asymmetric GRPO clipping with $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.3$, gradient-norm clipping at 1.0, and one policy update per sampled rollout group. Only the final three Transformer blocks, final RMSNorm layer, and language-model head are updated; all earlier layers and the embedding layer remain frozen. We use 500 protein representatives from the Prot2RNA training split, selected after clustering at 30% sequence identity to reduce redundancy, together with 50 validation representatives selected analogously. Each update samples a minibatch of 16 proteins with $G = 16$ candidate CDSs per protein, giving 256 rollouts per update. Candidates are generated with four iterative decoding steps, a cosine masking schedule, temperature 1.5, nucleus sampling $p = 0.95$, and synonymous decoding enforced throughout.

E Test-time reinforcement learning

Prot2RNA-TTRL further adapts a globally reinforcement-finetuned model to the synonymous design space of a single protein at test time. For NanoLuc, TTRL is initialized from the global Prot2RNA-RFT checkpoint trained without KL regularization ($\beta = 0$), which achieved the highest normalized-MFE reward among the global RFT variants. A configurable number of final Transformer blocks, together with the final RMSNorm layer and language-model head, are unfrozen for

adaptation. At each TTRL step, the current policy generates a group of synonymous CDS candidates for NanoLuc using stochastic iterative decoding; normalized MFE is computed for each candidate, rewards are standardized within the sampled group, and the policy is updated using the same GRPO objective as global RFT. Synonymous decoding is enforced throughout, and no additional KL regularization is used during TTRL. Each configuration is optimized for at most 50 GRPO update steps.

Because suitable test-time adaptation hyperparameters were not known a priori, we used a staged search. We first varied the learning rate over $\{10^{-4}, 10^{-5}, 10^{-6}\}$ and the number of unfrozen Transformer blocks over $\{1, 3, 6\}$, with $G = 16$, temperature $\tau = 1.5$, four decoding iterations, and a cosine masking schedule. We then varied $G \in \{8, 16, 32, 64\}$ and $\tau \in \{1.0, 1.5, 2.0\}$, followed by a final sweep over masking schedules and decoding iterations $\{1, 2, 4, 8, 12, 16\}$. The highest-reward sequence identified during this staged target-specific search was obtained with a learning rate of 10^{-4}, three unfrozen Transformer blocks, group size $G = 16$, temperature $\tau = 1.5$, two decoding iterations, and a square-root masking schedule. This sequence, which was selected for experimental evaluation, reached a normalized MFE of 0.4977. This search was performed for NanoLuc only; we did not evaluate TTRL on additional targets in this work.

F Dataset construction and splits

Protein–coding-sequence (CDS) pairs were extracted from NCBI RefSeq annotations of two human genome assemblies: GRCh38.p14 (accession GCF_000001405.40) and the telomere-to-telomere T2T-CHM13v2.0 assembly (accession GCF_009914755.1). Extracted pairs were subjected to quality control: verification of coding-sequence integrity, confirmation that each CDS translates to its annotated protein, removal of duplicate protein–CDS pairs, and exclusion of mitochondrial transcripts, yielding 118,224 protein–CDS pairs.

Cluster-level splitting. All partitions were constructed at the protein-cluster level using MMseqs2 (Steinegger & Söding, 2017; 2018), so that similar proteins and their CDSs never appear across more than one split. Clusters containing proteins selected as prospective future experimental targets, including NanoLuc, firefly luciferase, EPO, and several other candidate proteins, were identified at 60% sequence identity / 80% coverage and removed from the training pool before any train/validation/test partitioning, so that experimental validation (Section 3.4) is performed on proteins whose clusters were never seen during training. Test clusters were constructed first, at 60% identity / 80% coverage, from which 3,004 protein–CDS pairs were sampled. The remaining sequences were re-clustered at 80% identity / 80% coverage to construct a validation set of approximately 2,000 pairs; the remaining clusters were used for training.

G Sequence-property metrics: definitions

None of the metrics below alone is sufficient to predict protein expression; each captures a distinct aspect of coding-sequence composition, and none besides normalized MFE is used as an optimization target anywhere in this work.

Codon accuracy. Codon accuracy measures codon-level similarity between a generated CDS and the natural CDS references encoding the same protein. For each generated sequence, we calculate the fraction of codon positions matching each available natural reference and report the highest value across references. Codon accuracy therefore measures similarity to the closest natural CDS rather than correctness, since multiple synonymous sequences can validly encode the same protein.

Codon Adaptation Index (CAI). CAI measures the extent to which a coding sequence preferentially uses synonymous codons that are frequent in highly expressed human genes (Sharp & Li, 1987), computed using the EMBOSS implementation (version 6.6.0) with the human codon usage table `Ehuman.cut` (Rice et al., 2000).

GC content. The fraction of guanine and cytosine nucleotides in a coding sequence (Kudla et al., 2006), which affects RNA secondary-structure stability through the stronger base pairing of G–C compared to A–U pairs (Hia et al., 2019).

Uridine (U) fraction. The proportion of uridine nucleotides in a coding sequence, relevant to therapeutic mRNA design since synthetic mRNA therapeutics commonly substitute modified uridine analogues such as N^1-methylpseudouridine to reduce innate immune activation while improving transcript stability and protein expression (Karikó et al., 2008; Andries et al., 2015).

Normalized minimum free energy (normalized MFE). MFE is computed with the ViennaRNA package (version 2.7.2) (Lorenz et al., 2011). We report the length-normalized value

$$\text{Normalized MFE} = -\frac{\text{MFE}}{L},$$

where L is the coding-sequence length in nucleotides, so that larger normalized MFE corresponds to greater predicted structural stability. This is the only metric in this appendix used as an optimization objective in this work.

H Fréchet Transcript Distance: full methodology and calibration

Embedding and distance. Given an embedding function $f_{\text{LM}}(\cdot)$, natural and generated CDSs are mapped to sequence-level representations $R = \{f_{\text{LM}}(r)\}$ and $G = \{f_{\text{LM}}(r')\}$. Each population is approximated by a multivariate Gaussian with means μ_R, μ_G and covariances Σ_R, Σ_G, and FTD is the Fréchet distance between these Gaussian approximations,

$$\text{FTD}(R, G) = \|\mu_R - \mu_G\|_2^2 + \text{Tr}\left(\Sigma_R + \Sigma_G - 2(\Sigma_R\Sigma_G)^{1/2}\right), \tag{7}$$

by analogy with the Fréchet Inception Distance used for image generation. FTD depends on the choice of transcript embedding model; here we instantiate f_{LM} using the final `[CLS]` representation of Sanofi's CodonBERT (Li et al., 2024), pretrained on natural coding sequences. We report FTD as supporting evidence for population-level similarity to natural CDSs, alongside its finite-sample calibration below, rather than as a primary contribution.

Finite-sample calibration. Because FTD depends on sample estimates of the embedding mean and covariance, its value varies with population size even when both sequence sets are drawn from the same underlying distribution. Using 91,868 natural human CDSs whose lengths were within CodonBERT's 1,022-token input limit, repeatedly partitioned into two non-overlapping subsets of equal size n, mean FTD decreases from 30.16 ± 1.56 at $n = 50$ to 1.73 ± 0.04 at $n = 3{,}000$, approaching 0.52 ± 0.01 at $n = 10{,}000$ (Figure 8).

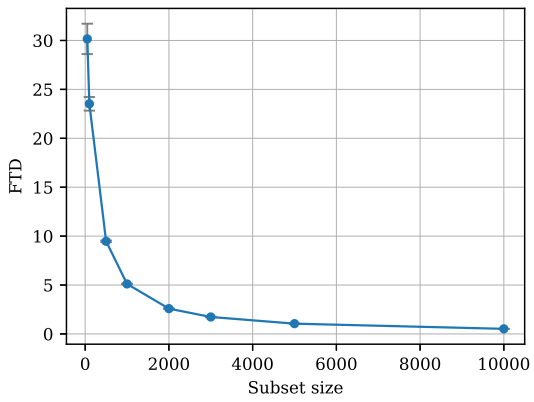

Figure 8: Sample-size dependence of the Fréchet Transcript Distance (FTD) noise floor: FTD computed between two non-overlapping random subsets of natural human CDSs, as a function of subset size n (10 random partitions per n).

Interpreting reported FTD values. For validation-set analyses ($n \approx 1{,}700$), the expected finite-sample baseline is approximately 3, whereas comparisons using the combined validation and test sets ($n \approx 4{,}700$) have a baseline of approximately 1.1–1.2. FTD values should therefore be compared with the appropriate sample-size baseline rather than interpreted relative to zero.

I TRAINING-OBJECTIVE VARIANTS: FULL DEFINITIONS

Section 3.1 compares four pretraining/finetuning combinations on the validation set; all variants use identical architecture, data splits, training budget, and optimization settings, with only the pretraining/finetuning objective changed.

NTP pretraining and finetuning. NTP pretraining uses a standard causal objective on protein and CDS sequences independently. During finetuning, the protein sequence is a conditioning prefix and the CDS is generated autoregressively.

DLM pretraining. For DLM→DLM, the diffusion masking objective is used during both pretraining and finetuning, with masking ratio $t \sim \mathcal{U}(0, 1)$ sampled independently per example (versus MLM's fixed 15%).

MLM finetuning. MLM finetuning uses the same conditional setup as DLM finetuning but retains the fixed 15% masking ratio; at generation time the CDS is reconstructed in a single unmasking step, creating a larger train/inference mismatch than DLM finetuning.

Effect of pretraining. Training the same architecture from random initialization with the DLM finetuning objective reaches mean codon accuracy 0.645 ± 0.191, compared with 0.688 ± 0.203 after MLM pretraining (Wilcoxon signed-rank $p = 5.3 \times 10^{-100}$, rank-biserial $r = 0.606$), showing that MLM pretraining improves conditional diffusion finetuning under the matched training setup.

J ROPE AND PADDING ABLATION: CONFIGURATION DETAILS

The three configurations compared in Table 2 are: (i) aligned RoPE + mid-padding (our default), where protein sequences are padded to the batch-maximum length before appending the CDS and $\phi(p_i) = \phi(r_i) = i$; (ii) standard RoPE + mid-padding, with the same padding but continuously increasing positional indices; and (iii) standard RoPE, no mid-padding, where the CDS immediately follows the unpadded protein with standard consecutive indices.

K FULL BASELINE RESULTS: ADDITIONAL DETAIL

Under unconstrained decoding (no synonymous-codon constraint), all generated test-set Prot2RNA sequences translate exactly to the target protein. Under the same unconstrained setting, CodonBERT (evaluated without its codon-correction post-processing step) shows amino-acid mismatches in 10.5% of sequences, while Trias shows mismatches in 0.4% of sequences.

L FULL RFT EVALUATION ACROSS THE KL SWEEP

Only normalized MFE is optimized during reinforcement finetuning. Figure 9 shows the resulting secondary changes in CAI, GC content, and uridine fraction across the full KL sweep, and Figure 10 the corresponding population-level shift measured by FTD. GC content increases as KL regularization is reduced, from a median of 0.575 for Prot2RNA to 0.673 at $\beta = 0$ (natural test median: 0.526); uridine fraction decreases correspondingly from 0.196 to approximately 0.15. CAI behaves non-monotonically: it initially increases from 0.822 for Prot2RNA to 0.843 at $\beta = 0.02$, then decreases as regularization is weakened, reaching 0.771 at $\beta = 0$ – close to the natural test-set median of 0.767. We interpret β as constraining how far the policy departs from the natural prior while optimizing the reward, rather than as controlling a single naturalness axis.

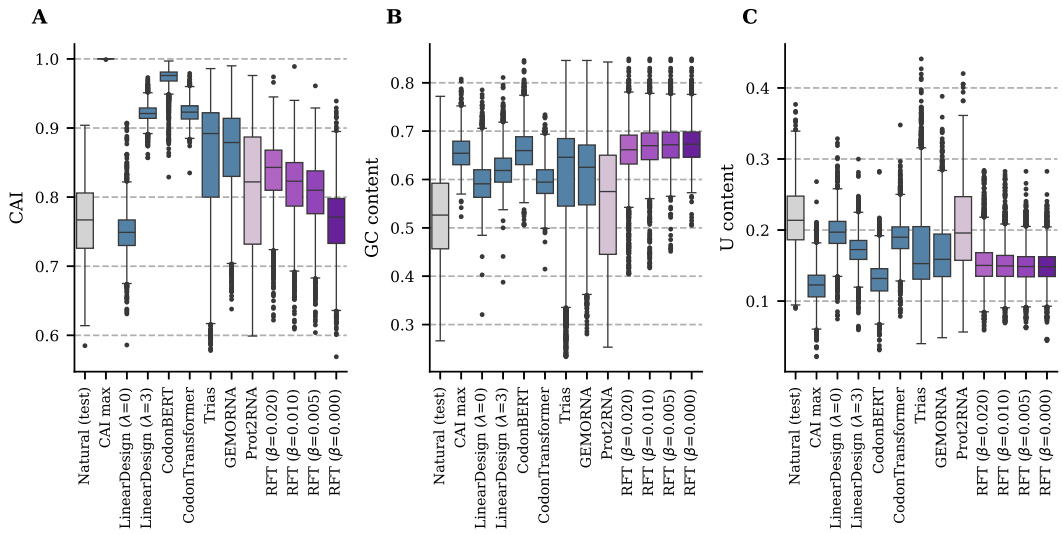

Figure 9: **Sequence-property distributions before and after RFT on test set.** (A) Codon Adaptation Index (CAI), (B) GC content, and (C) uridine fraction for natural test sequences, baseline methods, Prot2RNA, and RFT policies.

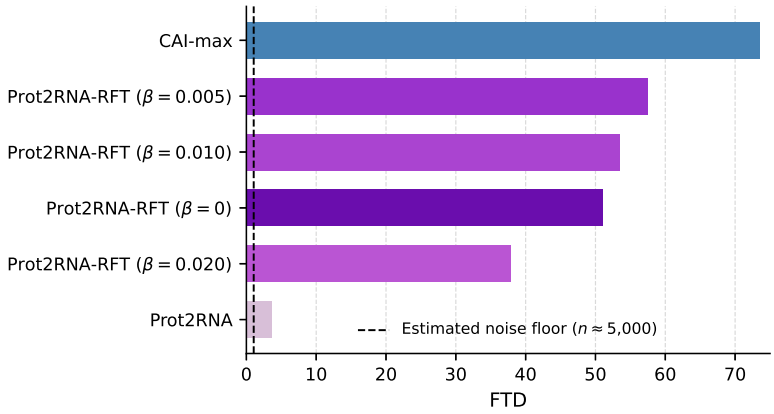

Figure 10: Fréchet Transcript Distance relative to natural human CDSs for CAI-max as a baseline, Prot2RNA, and RFT policies across KL coefficients ($n \approx 4{,}700$, combined val+test).

Diversity. Within-protein sequence diversity was quantified as the mean pairwise codon Hamming distance, normalized by sequence length, among 32 stochastic samples for each of 30 validation proteins (Figure 3B).

M SEED ROBUSTNESS OF RFT ($\beta = 0.005$): METHODOLOGY

To assess whether the reward shift from global RFT is specific to one training run, we trained three independent $\beta = 0.005$ policies from seeds 42, 43, and 44 and evaluated each on the held-out test set; results are reported in Table 4 in the main text. All three seeds converge to closely matching mean normalized MFE and comparable FTD, consistent with the reward shift being a property of the fine-tuning procedure rather than of a single training run. Per-seed distributions of secondary sequence properties (GC content, uridine fraction, CAI) and within-protein diversity were not separately collected beyond the two metrics reported in the main text.

N BEST-OF-N SAMPLING ANALYSIS: FULL METHODOLOGY

Figure 4 compares reward adaptation against increased sampling from the corresponding pre-adaptation policy. Panel A compares global RFT against best-of-N sampling from Prot2RNA across held-out proteins, while panel B compares NanoLuc-specific TTRL against best-of-N sampling from the global RFT policy used to initialize TTRL. We describe the two analyses below.

N.1 BEST-OF-N COMPARISON FOR GLOBAL REINFORCEMENT FINETUNING

For Figure 4A, we randomly chose 100 proteins from the test set and, for each, generated a pool of 256 stochastic CDS samples from Prot2RNA using the same iterative decoding configuration used for RFT evaluation (temperature 1.5, nucleus sampling $p = 0.95$). For each sampling budget $N \in \{1, 2, 4, 8, 16, 32\}$, we repeatedly drew N distinct candidates from this pool without replacement and recorded the maximum normalized MFE among the selected sequences. This was repeated 100 times per protein and per N, and the resulting maxima were averaged. For comparison, we generated 64 stochastic samples per protein from the $\beta = 0$ and $\beta = 0.005$ RFT policies using the same decoding configuration and report the mean normalized MFE across samples per protein, averaged across proteins. Mean best-of-N normalized MFE for Prot2RNA increases from 0.331 at $N = 1$ to 0.354 at $N = 32$, while ordinary stochastic sampling from the RFT policies reaches 0.468 ($\beta = 0.005$) and 0.482 ($\beta = 0$).

N.2 BEST-OF-N COMPARISON FOR NANOLUC TEST-TIME ADAPTATION

For Figure 4B, we generated a pool of $1,024$ stochastic CDS samples from the $\beta = 0$ Prot2RNA-RFT checkpoint used to initialize TTRL, using four decoding iterations, a cosine masking schedule, temperature 1.5, and nucleus sampling $p = 0.95$. For each sampling budget $N \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$, we repeatedly drew N distinct candidates from this pool without replacement and recorded the maximum normalized MFE among the selected sequences. This was repeated 100 times for each N, and the resulting maxima were averaged. For $N = 1{,}024$, we report the maximum normalized MFE over the full sample pool.

For comparison, we generated 64 stochastic samples from the final TTRL policy and report their mean normalized MFE. We additionally report the highest-reward sequence identified during the complete staged TTRL procedure described in Appendix E. This comparison evaluates whether increased sampling from the fixed starting RFT policy can recover the reward levels obtained after target-specific policy adaptation.

O NANOLUC EXPERIMENTAL VALIDATION: FULL METHODOLOGY

Construct design. All evaluated coding sequences encode the same NanoLuc luciferase protein and differ only in synonymous codon choice. The commercial Promega NanoLuc coding sequence was used as the reference. The panel additionally included a deliberately low-GC% synonymous control; Yellowstone, a published NanoLuc CDS from Leppek et al. (2022) that showed comparatively low protein output in their experiments and is included here as a negative control; designs from LinearDesign (Zhang et al., 2023) and GEMORNA (Zhang et al., 2024); and sequences generated by Prot2RNA, Prot2RNA-RFT, and Prot2RNA-TTRL.

The low-GC% control was constructed by preferentially selecting synonymous codons that reduce GC content, producing a sequence with low normalized MFE. For LinearDesign, we generated sequences across different values of its λ parameter and selected an intermediate design ($\lambda = 3.3$) rather than the most structure-maximizing setting. The GEMORNA sequence was generated using the published GEMORNA model and pipeline. Prot2RNA and the two global Prot2RNA-RFT sequences ($\beta = 0.005$ and $\beta = 0$) were generated using deterministic greedy decoding. Prot2RNA-TTRL was obtained through the test-time reinforcement-learning procedure described in Appendix E, with the experimentally tested construct selected as the highest-reward sequence identified during the staged search.

mRNA preparation. DNA fragments containing the respective NanoLuc coding sequences were synthesized and PCR-amplified to generate templates for in vitro transcription. Common 5' and

3' UTR sequences were introduced during PCR amplification. The 5' UTR was derived from α-globin/HBA1, while the 3' UTR contained an HBA1-derived sequence followed by an additional 30-nt sequence derived from the BNT COVID-19 3' UTR. The PCR products were purified and used for IVT. All mRNAs were prepared using the same construct context and contained a CleanCap cap, a poly$(A)_{50}$ tail, and complete substitution of uridine with N1-methylpseudouridine. Thus, the synonymous NanoLuc CDS was the variable of interest across constructs.

Cell culture and transfection. NanoLuc mRNA constructs were evaluated in HEK293T cells. Cells were seeded in 12-well plates and transfected one day later at approximately 70% confluency. Each well received 100 ng mRNA using 0.3 μL MessengerMAX reagent.

NanoLuc activity assay. NanoLuc activity was measured 72 h after transfection using the Nano-Glo Luciferase Assay System (Promega). Cells were lysed, and 10 μL of lysate from each replicate well was transferred to a white 96-well assay plate. Nano-Glo substrate (50 μL) was then added, and luminescence was measured using a Tecan plate reader. Two replicate wells were measured per construct within the same experiment. No additional normalization to cell number, protein concentration, viability, or a co-transfected reporter was performed. The mean luminescence across the two replicate wells was expressed relative to the Promega NanoLuc reference, which was assigned a relative output of 1.0.