ProMORNA: An End-to-End Generative Framework for Full-Length mRNA Design

Zixi Shao

Abstract

Designing therapeutic messenger RNA (mRNA) requires generating full-length transcripts that optimally balance intracellular stability and translation efficiency. While generative models have advanced mRNA design, existing approaches rely on region-wise generation, expensive inference-time search, or target-specific model updates. To overcome these limitations, we introduce ProMORNA, a zero-shot, end-to-end generative framework for full-length mRNA design. ProMORNA aligns the design process with modern Large Language Model training pipelines: it utilizes a BART-style architecture pre-trained on over 6 million natural protein–mRNA pairs, followed by a two-stage post-training process comprising cold-start supervised fine-tuning and multi-objective reinforcement learning. By standardizing and aggregating relative advantages at the individual metric level with domain-specific stabilization, our post-training stage directly aligns the generative policy with conflicting biological constraints while avoiding direct scalarization of raw objective scores. *In silico* evaluations demonstrate that ProMORNA pushes the Pareto frontier of predicted stability and translation efficiency beyond standard baselines and GEMORNA, a leading state-of-the-art method. In three independent *in vitro* Fluc2P experiments, Candidate 2 achieved a 12–72 h reporter-activity AUC 3.33 times that of the conventional Benchmark (Holm-adjusted $P = 0.01285$), while its estimated 1.29-fold advantage over the reported best GEMORNA candidate remained exploratory. These results support ProMORNA as a promising framework for generating functionally active full-length mRNA candidates.

Keywords: mRNA Design, Generative Biology, Large Language Models, Multi-Objective Optimization, Reinforcement Learning

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖公示

Contents

1 Introduction **3**

2 Methods **5**
2.1 Task Formulation and Model Architecture 5
2.2 Pre-training 5
2.3 Post-training 6
2.3.1 Cold-Start Supervised Fine-Tuning 6
2.3.2 Biological Objectives and Design Constraints 7
2.3.3 Multi-Objective Reinforcement Learning 8

3 Results **10**
3.1 *In silico* mRNA Generation Performance 10
3.2 *In Vitro* Validation 12

4 Conclusion **15**

References **16**

A Experimental Setup **19**
A.1 Model Training 19
A.2 *In Silico* Experimental Setup 20
A.3 *In Vitro* Validation 20

B Reinforcement Learning Dynamics **22**

C MO-GRPO Ablation **24**

D Additional mRNA Property Analysis **25**

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

1 Introduction

The success of messenger RNA (mRNA) vaccines against SARS-CoV-2 has highlighted the broad potential of mRNA-based medicines [1, 2, 3, 4]. Creating effective mRNA medicines requires computational design of mRNA sequences based on the amino acid sequence of the desired protein [5].

Full-length mRNA design can be viewed as a conditional generation and multi-objective optimization problem: given a protein sequence of amino acids, the goal is to generate an mRNA sequence of nucleotides that, (i) translates back into that protein, and (ii) optimizes for certain competing biological objectives. Specifically, a full-length mRNA sequence has a central coding sequence (CDS) bordered by untranslated regions (UTRs) at both ends. The CDS is divided into codons of three adjacent nucleotides, where each codon naturally translates into a certain amino acid determined by the genetic code. Thus, the number of codons in the CDS must be equal to the number of amino acids in the target protein, and each codon must be selected from a fixed set of synonymous codons that translate into the amino acid in the corresponding position. In contrast, the UTRs are largely unconstrained but work with the CDS to shape biomolecular dynamics and finally protein output [6, 7].

The primary challenge of mRNA design is the large design space, which makes exhaustive search infeasible. As illustrated in Fig. 1, the number of possible CDSs alone for a given protein grows exponentially with the protein length, not even considering the unconstrained compositions of the UTRs.

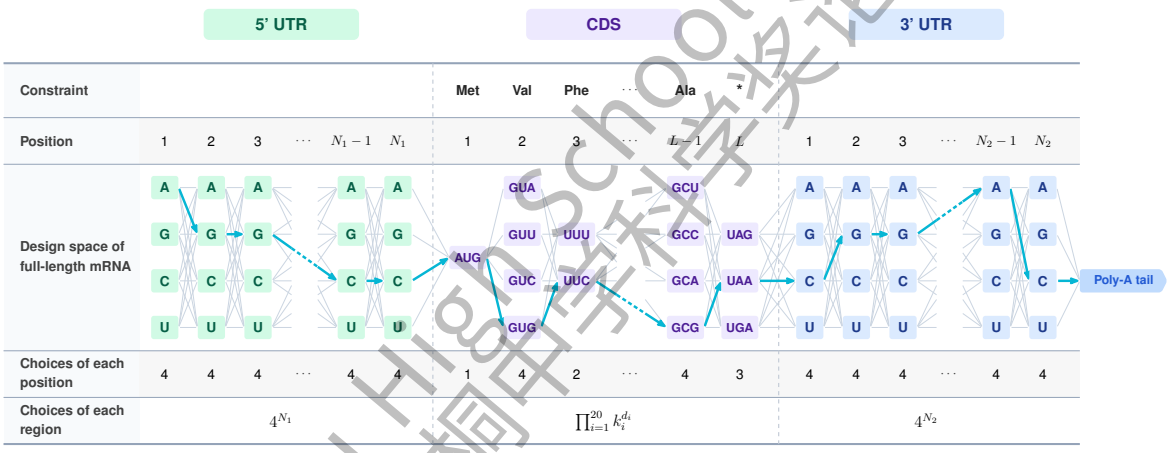


Figure 1: The vast potential mRNA search space of a target protein.

A primary objective in mRNA sequence design is to maximize functional protein expression, which is shaped by stability and translation efficiency [7, 8]. However, the complex biological mechanisms that affect these properties remain incompletely understood, making it difficult for rule-based algorithms to accurately predict and optimize for downstream biological properties. Conventional methods, such as LinearDesign [9], approach this challenge by jointly optimizing simple heuristic proxy metrics like Minimum Free Energy (MFE) and Codon Adaptation Index (CAI) [10]. Although these algorithms can achieve mathematical optimality within their defined search spaces, their real-world biological performance is fundamentally bounded by the limitations of these rigid, rule-based heuristics.

Recently, generative models have emerged to address limitations of traditional heuristic algorithms. For example, mRNAutilus [11] treated mRNA design as a discrete diffusion process guided by inference-time tree search; GEMORNA [12] used separate Transformer models to independently generate the 5' UTR, CDS, and 3' UTR. Most recently, mRNA-GPT [13] proposed jointly optimizing these regions through full-length sequential generation using a unidirectional, decoder-only language model.

However, these existing generative frameworks are not strictly end-to-end and are limited by their architecture. mRNAutilus requires a CDS template and two UTR lengths as inputs to initialize the

2026 Sci-T You High School Science Award
仅用于2026年度校园科学论文公示

masked diffusion process, which is constrained by hand-crafted priors; GEMORNA does not directly optimize for downstream biological objectives, which limits the ability to explore candidates beyond the training distribution, and concatenates outputs from three independent models, which may overlook long-range context-dependency between different regions [7, 14]; mRNA-GPT is partially or fully blind to the protein target when generating the 5' UTR and CDS regions, limiting its capacity to plan beyond the current amino acid.

Furthermore, these methods rely on computationally intensive inference-time scaling: mRNAutilus runs tree search at every diffusion step; GEMORNA requires wet-lab screening on a large number of generated candidates and optionally utilizes beam search for CDS generation; mRNA-GPT applies Iterative Supervised Fine-Tuning (SFT) or Reinforcement Learning (RL) for each specific target from a fixed base model, generating thousands of candidates per iteration per target. The vast resources required for these models at inference time limit their application to practical scenarios such as personalized mRNA vaccines [15, 16], where rapid computational turnaround is critically required.

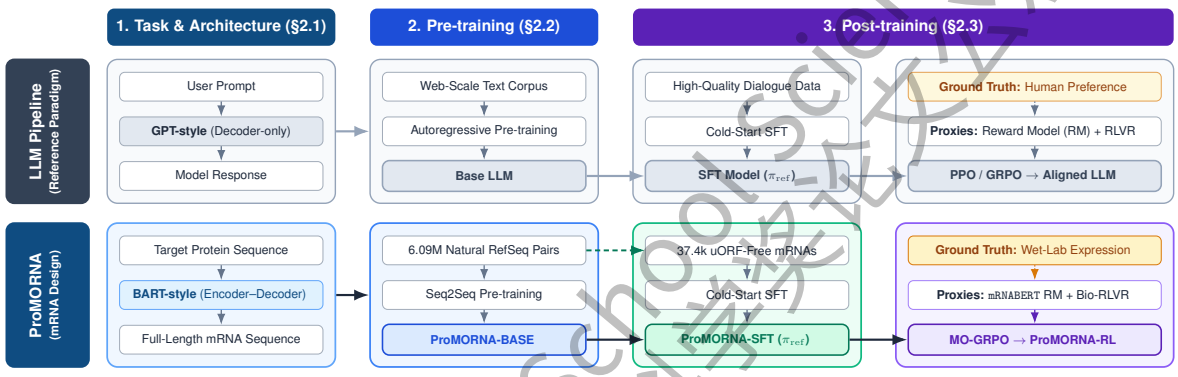


Figure 2: Overview of the ProMORNA framework and its conceptual alignment with modern LLM development.

To overcome these challenges, we propose **ProMORNA** (**Pro**tein-conditioned **M**ulti-**O**bjective **RNA** generation), an end-to-end generative framework for full-length mRNA design (Fig. 2). First, we pre-train a BART-style encoder–decoder generative model on public data. Next, inspired by modern LLM alignment [17, 18, 19, 20], we employ a two-stage post-training pipeline consisting of cold-start supervised fine-tuning (SFT) followed by multi-objective reinforcement learning (RL) to optimize for competing biological metrics. *In silico* experiments illustrate that ProMORNA improves the Pareto frontier for predicted stability and translation efficiency relative to standard supervised baselines and the existing generative mRNA design method, GEMORNA [12]. Wet-lab evaluation of seven Pareto-selected candidates on Fluc2P provides functional evidence for the generated designs.

We summarize the core contributions of our framework as follows:

- **Protein-conditioned end-to-end mRNA generation:** We introduce an encoder–decoder framework that generates complete mRNA transcripts, including the 5' UTR, CDS, and 3' UTR, directly from a target protein sequence, without requiring additional inputs at inference time. The model is given maximum freedom in generation and optimization, and only a minimal set of basic biological rules is enforced throughout the whole pipeline.
- **Multi-stage, target-agnostic model alignment:** We structure full-length mRNA design into a progressive pre-training and two-stage post-training pipeline. Our target-agnostic RL post-training generalizes to unseen proteins, enabling rapid, high-quality zero-shot generation without requiring target-specific model updates. Furthermore, the autoregressive model natively supports inference-time scaling, allowing users to optionally generate and filter large candidate pools to further explore the Pareto frontier when computational budget permits.

2026 S. T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

- **Pushing the Pareto frontier *in silico*:** We evaluate ProMORNA *in silico* against existing design strategies and assess generated candidates across translation efficiency, stability, normalized Minimum Free Energy (MFE), uridine content (U-content), and UTR length plausibility. Notably, we find that ProMORNA significantly improves the performance over supervised baselines and GEMORNA on the two primary predicted objectives: predicted stability and predicted translation efficiency.
- **Prospective biological validation:** We synthesize and evaluate seven Pareto-selected ProMORNA candidates alongside a conventional mRNA Benchmark and the GEMORNA design GMR-FL-F5 [12]. Three independent experiments identify candidates with higher time-integrated reporter activity than Benchmark.

2 Methods

In this section, we detail the **ProMORNA** framework illustrated in Fig. 2. Following the paradigm of modern large language model (LLM) development, we organize our methodology into three parts: (i) task formulation and model architecture (Section 2.1); (ii) large-scale evolutionary pre-training (Section 2.2); and (iii) two-stage post-training (Section 2.3), which comprises a cold-start supervised fine-tuning (SFT) stage followed by multi-objective reinforcement learning (RL).

2.1 Task Formulation and Model Architecture

Task Formulation. We formulate full-length mRNA design as a conditional sequence-to-sequence generation task. Given the amino acid sequence of a target protein $\mathbf{x}$, the goal is to generate a complete mRNA transcript $y = (y_1, \ldots, y_T)$ spanning the 5' UTR, CDS, and 3' UTR by modeling the conditional probability distribution $P_\theta(y \mid \mathbf{x}) = \prod_{t=1}^{T} P_\theta(y_t \mid y_{<t}, \mathbf{x})$, such that the CDS translates back into $\mathbf{x}$ while the full transcript optimizes downstream biological properties.

Model Architecture. Our model uses an encoder–decoder architecture based on the BART-small configuration [21] with a context limit of 1,024 tokens applied to both the encoder and the decoder, comprising 45.2M trainable parameters in total. As shown in Fig. 3, the encoder processes the input amino acid sequence of the target protein, while the decoder autoregressively generates the corresponding full-length mRNA sequence step by step. Because the encoder processes the full protein sequence before decoding begins, the decoder can condition every generated mRNA token on the complete target protein.

Tokenization. We apply per-amino-acid tokenization for the source protein sequences. For target mRNA sequences, we use a hybrid vocabulary: per-nucleotide tokenization for both UTRs, and per-codon tokenization for the CDS. Special separator tokens are inserted to mark region boundaries and prompt the model to switch between nucleotide-level and codon-level vocabularies.

2.2 Pre-training

Training Dataset. We curated vertebrate transcripts from NCBI RefSeq [22], prioritizing shorter sequences to ensure computational tractability and to remain within a practical length regime for future *in vitro* synthesis. We filtered out sequences labeled as "pseudo" and retained transcripts with canonical start and stop codons, a verified N-terminal methionine, and minimum 5' and 3' UTR lengths of 20 nt and 30 nt, respectively. Additionally, all retained sequences were required to fit within our 1,024-token limit under our hybrid tokenization scheme. This pipeline yielded 6,089,519 natural protein–mRNA pairs. From these data, the model learns the conditional distribution $P(y \mid \mathbf{x})$ of a full-length mRNA

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖公示

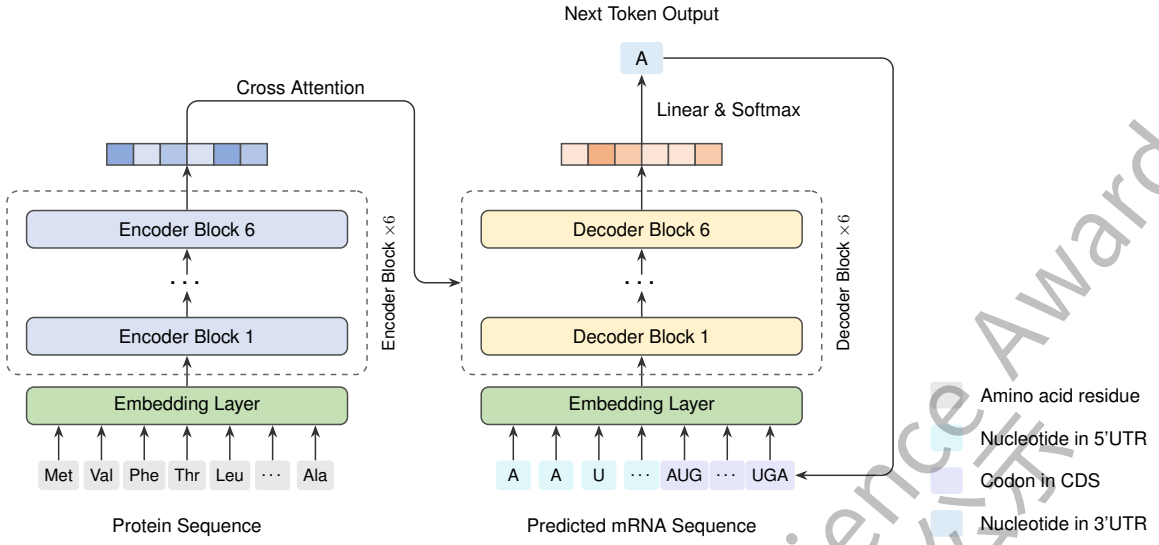


Figure 3: Modeling protein-to-mRNA with BART.

transcript y given a target protein sequence $\mathbf{x}$, implicitly capturing natural vertebrate codon usage biases and plausible UTR compositions.

Training Objective. The model is pre-trained by minimizing the standard sequence-to-sequence negative log-likelihood:

$$\mathcal{L}(\theta) = -\frac{1}{T}\sum_{t=1}^{T} \log P_\theta(y_t \mid y_{<t}, \mathbf{x}), \tag{1}$$

where θ represents the trainable parameters of the model, T is the total length of the target mRNA sequence, y_t is the ground-truth target token at time step t, $y_{<t}$ denotes the sequence of preceding target tokens, and $\mathbf{x}$ is the tokenized source input amino acid sequence. We denote the resulting pre-trained checkpoint as **ProMORNA-BASE** (or **BASE**).

2.3 Post-training

While large-scale pre-training captures natural evolutionary sequence grammar, designing therapeutic mRNAs requires shifting the model's generative prior toward high protein expression and explicitly optimizing competing biochemical properties. We therefore employ a two-stage post-training pipeline consisting of a cold-start supervised fine-tuning (SFT) stage (Section 2.3.1) followed by online reinforcement learning (Sections 2.3.2 and 2.3.3).

2.3.1 Cold-Start Supervised Fine-Tuning

Aligning with modern LLM post-training paradigms [17, 20], we employ SFT to establish a high-quality cold start for reinforcement learning. While pre-training captures broad natural evolutionary distributions, therapeutic mRNAs require maximized protein expression. Therefore, we curate a purified subset of 37,369 NM-prefixed RefSeq mRNAs strictly lacking upstream open reading frames (uORFs), which are natural regulatory elements that suppress translation [23]. Fine-tuning the base model on this subset using the objective in Eq. (1) shifts its generative prior toward therapeutic efficacy, yielding the **ProMORNA-SFT** (or **SFT**) checkpoint. Crucially, this step provides a biologically aligned reference policy π_{ref} for the subsequent RL phase, ensuring that the KL-divergence penalty does not force the model to mimic undesirable natural traits.

2026 S. T. Yau High School Science Award
仅用于2026丘成桐科学奖论文公示

2.3.2 Biological Objectives and Design Constraints

To guide policy optimization toward high-performing therapeutic candidates, we define $K = 5$ biological objectives that jointly capture functional efficacy, structural thermodynamics, sequence composition, and length regularity. For the i-th generated mRNA sequence, let $r_{i,k}$ denote its raw score on the k-th objective:

Predicted Stability ($k = 1$). Intracellular mRNA stability governs resistance to enzymatic degradation and determines the functional half-life of the transcript [7, 8]. Because stability cannot be computed analytically from primary sequence alone, we estimate $r_{i,1}$ using a ridge-regression head f_{S} trained on global-average-pooled sequence embeddings from a frozen mRNABERT encoder [24] (Fig. 4; see Appendix A.1 for training details). This primary functional objective is maximized.

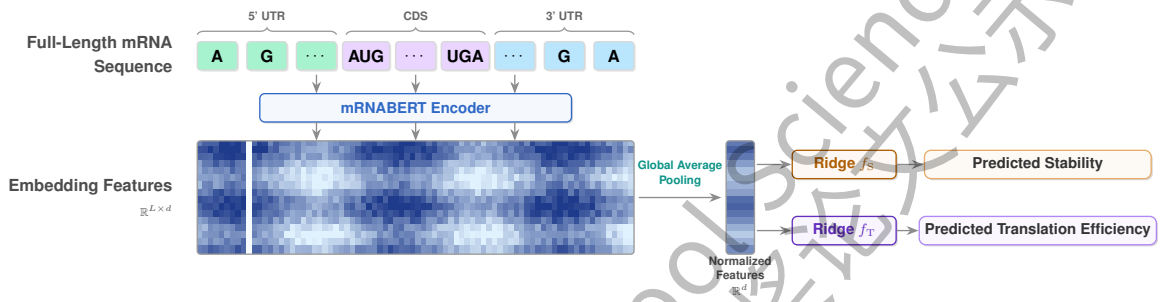


Figure 4: Predicting mRNA stability and translation efficiency with mRNABERT.

Predicted Translation Efficiency ($k = 2$). Translation efficiency measures the rate at which ribosomes initiate and elongate along the transcript to synthesize the target protein. Analogous to stability, we predict $r_{i,2}$ using an independent ridge-regression head f_{T} operating on the same pooled mRNABERT representation (Fig. 4). This second primary functional objective is also maximized.

Normalized Minimum Free Energy (MFE, $k = 3$). Secondary structure folding critically modulates both physical transcript longevity and ribosomal transit [8, 9]. We compute the full-transcript folding free energy using LinearFold [25] and normalize it by the total nucleotide length to obtain a length-invariant proxy for global thermodynamic stability. Since more stable secondary structures correspond to more negative free energy values, this objective is minimized.

U-content ($k = 4$). U-content is defined as the fractional proportion of uridine nucleotides across the full-length transcript. Because AU-rich elements accelerate mRNA decay [26] and uridine-rich motifs trigger innate immune receptors that suppress translation [27, 28], we explicitly minimize global uridine fraction.

UTR Length Plausibility ($k = 5$). To prevent length degeneration during unconstrained UTR generation, we regularize the generated 5' and 3' UTR lengths using Gaussian-shaped priors derived from natural transcripts, maximizing:

$$r_{i,5} = \sum_{j \in \{5,3\}} \exp\left(-\frac{1}{2}\left(\frac{u_{i,j} - \mu_{i,j}}{\sigma_j}\right)^2\right), \tag{2}$$

where $u_{i,j}$ denotes the nucleotide length of the generated j' UTR for sequence i, $\mu_{i,j}$ is the corresponding wild-type UTR length of the prompt protein in the training dataset, and σ_j is the global standard

2026 S.T. Yau High School Science Award 仅用于2026丘成桐中学科学奖公示

deviation of j' UTR lengths across the dataset. This term serves solely as a training-time regularizer, requiring no wild-type length priors during zero-shot generation or evaluation.

Generation Constraints. To guarantee sequence validity throughout rollouts, we combine prefix-level constrained decoding with a transcript-level post-hoc validity check. During autoregressive decoding, token sampling within the CDS is masked to allow only synonymous codons of the corresponding target amino acid, whereas UTR decoding is restricted to standard nucleotide tokens and region separators. Because prefix constraints alone cannot prevent premature truncation at the 1,024-token context limit or enforce global structural rules, each completed rollout undergoes a post-hoc verification of region boundaries, minimum UTR lengths, canonical start and stop codons, reading-frame integrity, and exact translation into the target protein. Moreover, following LinearDesign [9], transcripts containing contiguous double-stranded regions of 33 or more base pairs are rejected to mitigate double-stranded RNA-mediated innate immune responses. Sequences failing any post-hoc check bypass predictor evaluation, are excluded from group mean and standard deviation calculations, and receive a deterministic maximal penalty during advantage assignment.

2.3.3 Multi-Objective Reinforcement Learning

While standard RL algorithms collapse multiple evaluation metrics into a single explicit scalar reward, this static scalarization can be unstable for biological sequence design: biological metrics have different scales, and their variances shift asynchronously as the policy evolves. Recent studies in multi-objective LLM alignment, specifically Multi-Objective GRPO (MO-GRPO) [29] and Group reward-Decoupled Normalization Policy Optimization (GDPO) [30], have shown that applying vanilla GRPO [18] to summed rewards biases optimization toward high-variance objectives and causes reward signal collapse. Both methods address this by standardizing each objective independently within a rollout group prior to aggregation.

In ProMORNA, we adopt this decoupled group-relative standardization framework (retaining the name MO-GRPO, Fig. 5) and introduce several domain-specific improvements to address practical instabilities in biological sequence optimization.

L_2-Normalized Weighting and Dynamic Variance Damping. For a given prompt, the policy samples a group of G complete mRNA sequences. Let μ_k and σ_k be the empirical mean and standard deviation of the k-th objective scores across valid sequences within the current group. Extending the unweighted per-objective standardization in MO-GRPO [29], the aggregated baseline advantage A_i for sequence i is formulated as:

$$A_i = \frac{1}{\sqrt{\sum_{k=1}^{K} w_k^2}} \sum_{k=1}^{K} w_k \cdot \mathrm{clip}\left(\frac{r_{i,k} - \mu_k}{\sigma_k + \delta_k}, -c, c \right), \tag{3}$$

where w_k denotes the relative importance and optimization direction of the k-th objective ($w_k > 0$ for maximization and $w_k < 0$ for minimization). Normalizing the weights by their L_2 norm mathematically guarantees that, assuming the objectives are mutually orthogonal, the standard deviation of the aggregated advantages $\mathrm{std}(\{A_1, \ldots, A_G\})$ remains approximately 1. Crucially, whereas prior MO-GRPO [29] divides directly by σ_k, we introduce a dynamic damping factor $\delta_k > 0$ for each objective. When the group variance σ_k drops significantly ($\sigma_k \ll \delta_k$), δ_k smoothly pushes the normalized advantage contribution of this metric toward zero. This naturally prevents numerical instability and shifts the model's exploration focus toward other under-optimized metrics.

2026 S. T. Yau High School Science Award
仅用于2026丘成桐中学科学奖公示

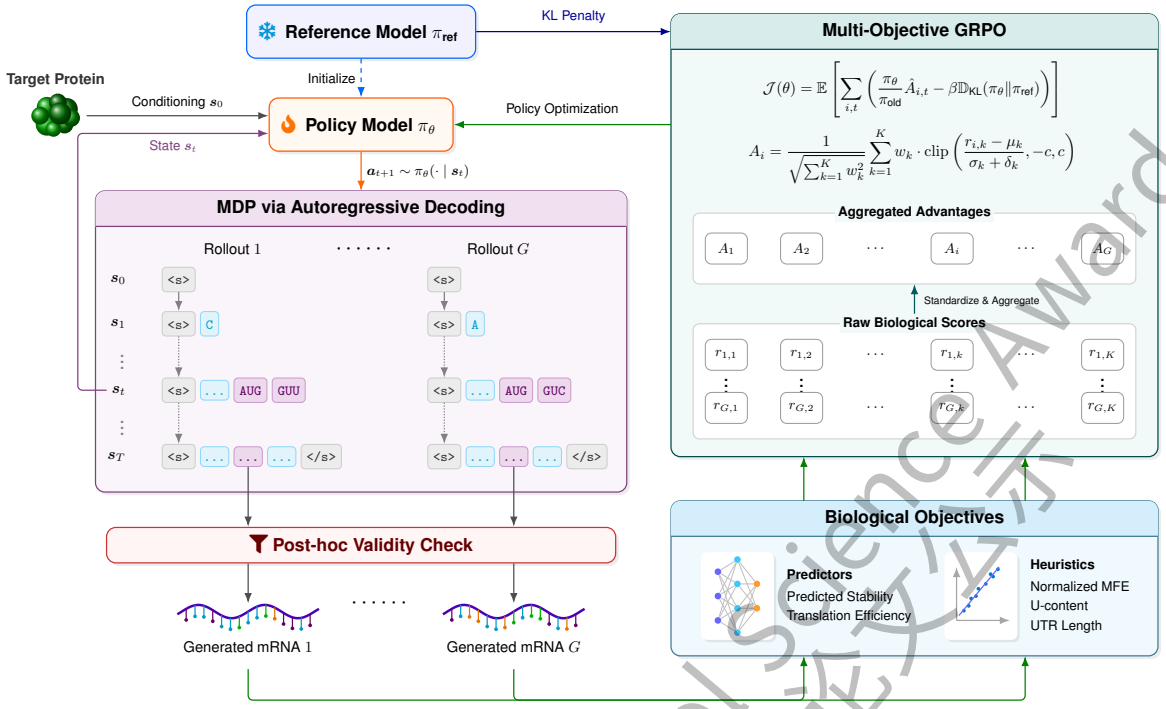


Figure 5: ProMORNA's post-training framework. `<s>` is the start token and `</s>` is the end token. Note that this is only a demonstration of the rollout process; in practice, mRNA sequences of varying lengths can be generated for the same target protein.

Component-Wise Extreme-Value Clipping. Unlike simple, rule-based NLP rewards, biological reward components are computed from heuristic or learned proxy predictors (`mRNABERT` and `LinearFold`), whose outputs on out-of-distribution rollouts may occasionally contain large deviations. In contrast to prior formulations [29, 30], we apply component-wise extreme-value clipping with bound $c = 3.0$ (Eq. (3)) prior to aggregation to limit the influence of outlier objective values.

Mean-Only Post-Centralization and Validity-Masked Penalty Injection. Because clipping, damping, and invalid-sequence penalties can shift the group-level mean away from zero, we apply a post-centralization step to obtain the centered sequence-level advantage $\hat{A}_i$, which is then broadcast across all tokens within the i-th sequence to serve as the token-level advantage $\hat{A}_{i,t}$:

$$\hat{A}_{i,t} = \hat{A}_i = A_i - \frac{1}{G}\sum_{j=1}^{G} A_j, \quad \forall t \in \{1, \ldots, |\boldsymbol{a}_i|\}. \tag{4}$$

Importantly, unlike GDPO [30], which re-divides the aggregated advantage sum by the batch standard deviation, we do not re-divide by the standard deviation in Eq. (4). This preserves the adaptive magnitude derived from the damping factor δ_k, allowing the advantage to decay naturally as the policy converges. For invalid sequences, we exclude them from μ_k and σ_k calculation and directly assign a maximal penalty $A_i = -c\sum|w_k|/\sqrt{\sum w_k^2}$ before post-centralization, thereby incorporating invalid samples into the policy update.

2026 S. T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

Token-Level Policy Gradient and Unbiased KL Regularization. Finally, the policy π_θ is updated to maximize:

$$\mathcal{J}(\theta) = \mathbb{E}_{\substack{\boldsymbol{s}_0 \sim P(S_0) \\ \{\boldsymbol{a}_i\}_{i=1}^{G} \sim \pi_{\text{old}}(\cdot|\boldsymbol{s}_0)}} \left[\frac{1}{\sum_{i=1}^{G} |\boldsymbol{a}_i|} \sum_{i=1}^{G} \sum_{t=1}^{|\boldsymbol{a}_i|} \left(\frac{\pi_\theta(\boldsymbol{a}_{i,t}|\boldsymbol{s}_{i,t})}{\pi_{\text{old}}(\boldsymbol{a}_{i,t}|\boldsymbol{s}_{i,t})} \hat{A}_{i,t} - \beta \, \mathbb{D}_{\text{KL}}\left[\pi_\theta \| \pi_{\text{ref}}\right] \right) \right], \tag{5}$$

where the KL divergence is approximated analytically as

$$\mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) = \frac{\pi_\theta(\boldsymbol{a}_{i,t}|\boldsymbol{s}_{i,t})}{\pi_{\text{old}}(\boldsymbol{a}_{i,t}|\boldsymbol{s}_{i,t})} \left(\frac{\pi_{\text{ref}}(\boldsymbol{a}_{i,t}|\boldsymbol{s}_{i,t})}{\pi_\theta(\boldsymbol{a}_{i,t}|\boldsymbol{s}_{i,t})} - \log \frac{\pi_{\text{ref}}(\boldsymbol{a}_{i,t}|\boldsymbol{s}_{i,t})}{\pi_\theta(\boldsymbol{a}_{i,t}|\boldsymbol{s}_{i,t})} - 1 \right). \tag{6}$$

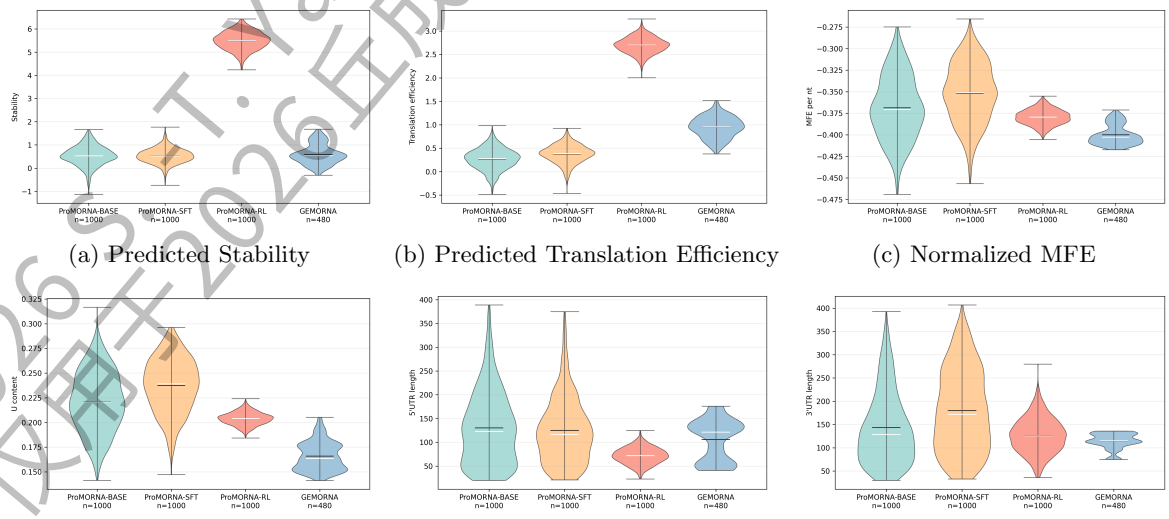
Here, different from vanilla GRPO [18], we (i) use a token-level policy gradient loss [31], assigning equal weights to every generated token; (ii) estimate the KL penalty with an importance-weighted estimator [32], which eliminates bias introduced by evaluating samples drawn from the old policy.

3 Results

This section presents *in silico* and *in vitro* results of ProMORNA on the widely used Fluc2P target. We compare ProMORNA with GEMORNA and supervised baselines to evaluate its mRNA generation performance. Appendix A presents the experimental setup, including the target protein, baselines, parameters, generation protocol, and the reward predictor training. Appendix B presents RL learning curves. Ablation studies demonstrate that standard scalarized reward optimization suffers from rapid reward hacking and KL-collapse, underscoring the necessity of our metric-level advantage formulation during RL post-training (Appendix C). Our additional analyses (see Appendix D) demonstrate that ProMORNA automatically learned several known biological rules, such as optimized GC content and structures more conducive to regulating protein expression. These findings may further indicate that the model enhanced its understanding of sequence structure in the absence of explicit supervisory signals.

3.1 *In silico* mRNA Generation Performance

In this subsection, we evaluate ProMORNA on Fluc2P using the reward-related metrics. We compare ProMORNA with GEMORNA and the two preceding checkpoints from our training pipeline, BASE and SFT.



(a) Predicted Stability (b) Predicted Translation Efficiency (c) Normalized MFE

(d) U-content (e) 5' UTR length (f) 3' UTR length

Figure 6: Comparison of three ProMORNA checkpoints and GEMORNA on reward-related metrics.

2026 ICLR Yau Hsin School Science Award 优秀2026丘成桐中学科学奖论文公示

The comparison results and ablation study are presented in Fig. 6. As shown in Figs. 6a and 6b, ProMORNA scores highest on the two primary functional objectives: predicted stability and translation efficiency. This indicates that our RL post-training effectively guides the generator toward candidates with improved functional proxy scores. While GEMORNA also outperforms the purely supervised BASE and SFT models, consistent with its strong region-wise mRNA generation design, ProMORNA builds on this by directly optimizing the full-length transcript using multi-objective reinforcement learning, rather than recombining independently generated UTR and CDS elements.

Figures 6c and 6d show that ProMORNA scores slightly higher than GEMORNA scores on normalized full-length MFE and U-content, but lower than BASE and SFT scores. Because mRNA expression relies on complex interactions among regional structure, codon usage, UTR context, innate immune sensing, and degradation, these secondary metrics should be evaluated alongside the primary functional objectives [5]. ProMORNA's moderate scores on these subsidiary metrics reflect the core goal of our multi-objective post-training: prioritizing a balanced, full-length trade-off over maximizing individual proxy metrics independently.

Regarding UTR length distributions (Figs. 6e and 6f), ProMORNA tends to generate shorter 5' UTRs and longer 3' UTRs compared to the supervised baselines. Since UTR length effects are highly context-dependent [14], this distribution suggests that the length-plausibility reward and functional objectives work together to guide ProMORNA toward an optimal, target-specific UTR length for Fluc2P instead of defaulting to a universally fixed length.

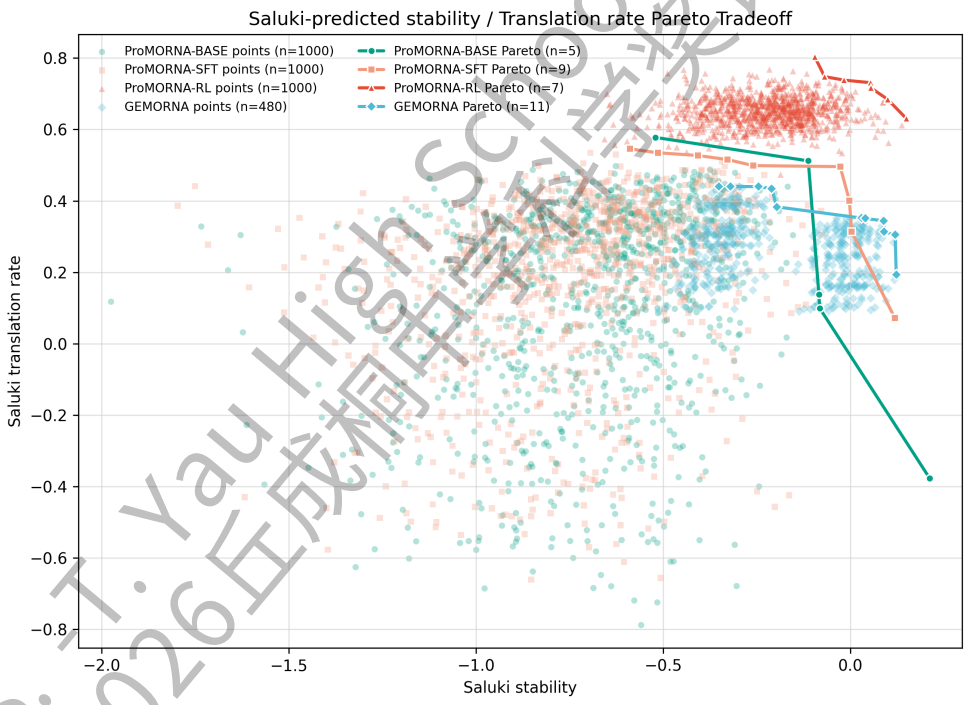


Figure 7: Orthogonal evaluation of Pareto frontiers for predicted stability and translation rate across four models using the independent evaluator Saluki [33].

Because the stability and translation efficiency scores in Figs. 6a and 6b are derived from the same mRNABERT-based proxy predictors used during RL training, optimizing directly against them could raise concerns of proxy overfitting. To provide an independent, orthogonal evaluation, we further assess all generated candidates using Saluki [33], an external sequence-based deep learning model that was entirely withheld from our training pipeline. Figure 7 illustrates the resulting Saluki-predicted stability and translation rate distributions alongside their corresponding Pareto frontiers across three ProMORNA models and GEMORNA. Consistent with our internal metrics, ProMORNA-RL achieves a pronounced upper-right shift of the Pareto frontier relative to both supervised baselines and GEMORNA. This cross-

2026 S.T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公开

evaluator consistency confirms that our multi-objective RL post-training captures genuine sequence-level determinants of stability and translation rather than exploiting artifacts of the training reward predictors, providing a reliable basis for candidate prioritization in subsequent wet-lab validation.

3.2 *In Vitro* Validation

Building on the orthogonal *in silico* evaluation, we select all seven Pareto-optimal candidates identified by Saluki from the 1,000 ProMORNA-RL sequences (Fig. 7) for prospective wet-lab validation. The candidates are tested alongside a conventional Benchmark and the leading GEMORNA design GMR-FL-F5 [12] in three independent Fluc2P experiments at 12, 24, 36, 48, and 72 h after transfection. Consistent with the GEMORNA study, the Benchmark construct comprises the pGL4.11 CDS flanked by the 5' and 3' UTRs of BNT162b2 [1].

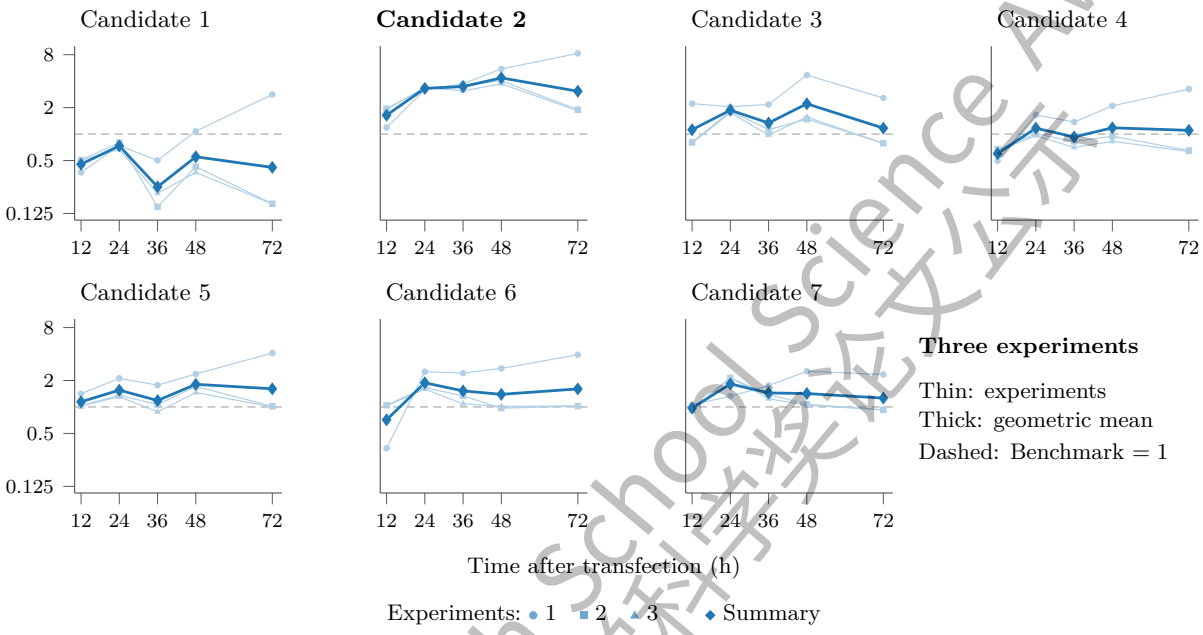
Figure 8a illustrates the Fluc activity normalized to the benchmark at each time point and Fig. 8b shows the AUC ratios of each candidate to Benchmark. Candidates 2 and 7 showed higher AUCs after seven-candidate Holm correction: Candidate 2 had a geometric-mean ratio of 3.33 (95% CI, 2.67–4.15; $P_{\mathrm{Holm}} = 0.01285$), and Candidate 7 had a ratio of 1.59 (1.43–1.77; $P_{\mathrm{Holm}} = 0.01767$). Candidate 2 is thus selected for closer comparison with the controls.

Figure 9a shows the raw activity trajectories of Candidate 2 and the two controls. Candidate 2 is numerically above both controls at every sampled time point in each experiment, peaking at 24 h and declining by 72 h. Using the same experiments, its AUC ratio relative to GMR-FL-F5 was 1.29 (95% CI, 1.09–1.52; unadjusted $P = 0.02319$; $P_{\mathrm{Holm}} = 0.09294$; Fig. 9b). This estimated 28.7% gain remains exploratory after correction across the seven GMR-FL-F5 comparisons.

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖公示

(a) Time courses of the seven candidates

Fluc activity / matched Benchmark (fold; log scale)



(b) AUC screening against Benchmark

Candidate	AUC ratio	Geometric mean [95% CI]	P (raw)	P_{Holm}
Candidate 1		0.53 [0.29, 0.98]	0.04770	0.23852
Candidate 2		3.33 [2.67, 4.15]	0.00184	0.01285
Candidate 3		1.70 [0.69, 4.16]	0.12559	0.50238
Candidate 4		1.04 [0.43, 2.51]	0.85488	0.85488
Candidate 5		1.44 [0.70, 2.96]	0.15964	0.50238
Candidate 6		1.61 [0.68, 3.79]	0.14077	0.50238
Candidate 7		1.59 [1.43, 1.77]	0.00295	0.01767

Figure 8: Screening ProMORNA candidates against Benchmark in three independent Fluc2P experiments.

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

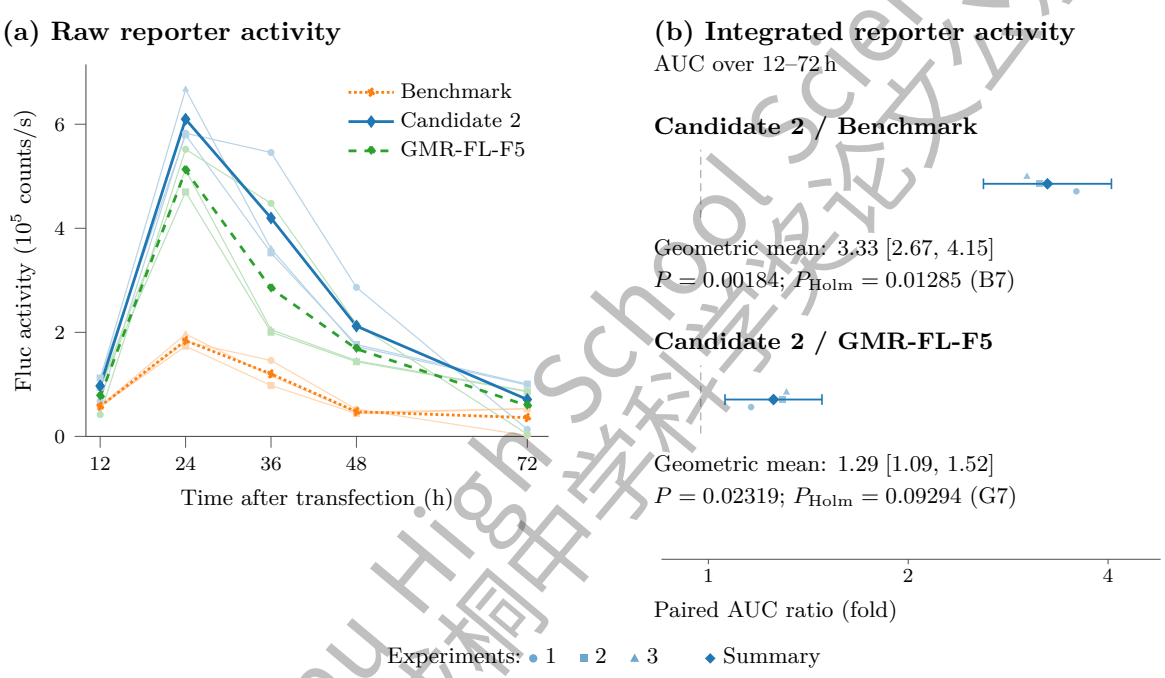


Figure 9: Candidate 2 compared with Benchmark and GMR-FL-F5 in the same three experiments. Holm-adjusted P values are shown for the corresponding seven-candidate families: B7 for Benchmark and G7 for GMR-FL-F5.

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

4 Conclusion

In this paper, we propose **ProMORNA**, an end-to-end generative framework for full-length mRNA design. ProMORNA generates complete mRNA transcripts, including the 5' UTR, CDS, and 3' UTR, directly from a target protein sequence. Using a BART-style encoder–decoder architecture, the model conditions every generated region on the full protein context, enabling joint design of coding and non-coding regions without requiring template inputs or updating model weights at inference time.

Following a progressive training pipeline that combines large-scale evolutionary pre-training with two-stage post-training, ProMORNA aligns its generative distribution with therapeutic mRNA design requirements. Unlike conventional scalar-reward optimization, our RL post-training computes and stabilizes sequence advantages at the level of individual biological metrics prior to aggregation. This formulation mitigates the instability caused by objectives with heterogeneous scales and shifting variances, providing a flexible mechanism for balancing competing design constraints.

In silico experiments on Fluc2P demonstrate that ProMORNA improves generated mRNA candidates over supervised baselines and GEMORNA, particularly on the two primary predicted objectives: predicted stability and predicted translation efficiency. Moreover, Pareto frontier analysis confirms that our post-training optimizes the model toward generating mRNA candidates with better multi-objective trade-offs. In the *in vitro* screen, Candidates 2 and 7 show significantly higher time-integrated reporter activity than Benchmark after Holm adjustment. These results demonstrate that ProMORNA is capable of generating functionally active full-length mRNA candidates.

A primary limitation of the current study is that our prospective biological validation was restricted to seven Pareto-selected candidates for a single reporter protein, Fluc2P. The retrospective AUC analysis and prioritization of Candidate 2 require independent validation, particularly against GMR-FL-F5. Reporter activity does not directly measure mRNA stability or translation efficiency. Extensive *in vitro* experiments with longer transcripts across diverse therapeutic contexts are required to rigorously establish the generalizability and practical applicability of ProMORNA. This future work will provide a stronger foundation for applying ProMORNA to therapeutic mRNA design. Moreover, improving the accuracy and experimental robustness of the stability and translation efficiency predictors will be essential for more reliable reward modeling.

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

References

[1] F. P. Polack et al., "Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine," *New England Journal of Medicine*, vol. 383, no. 27, pp. 2603–2615, 2020. DOI: 10.1056/NEJMoa2034577

[2] L. R. Baden et al., "Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine," *New England Journal of Medicine*, vol. 384, no. 5, pp. 403–416, 2021. DOI: 10.1056/NEJMoa2035389

[3] N. Pardi, M. J. Hogan, F. W. Porter, and D. Weissman, "mRNA vaccines—a new era in vaccinology," *Nature Reviews Drug Discovery*, vol. 17, no. 4, pp. 261–279, 2018. DOI: 10.1038/nrd.2017.243

[4] S. Qin et al., "mRNA-based therapeutics: Powerful and versatile tools to combat diseases," *Signal Transduction and Targeted Therapy*, vol. 7, no. 1, 2022, Art. no. 166. DOI: 10.1038/s41392-022-01007-w

[5] M. Metkar, C. S. Pepin, and M. J. Moore, "Tailor made: The art of therapeutic mRNA design," *Nature Reviews Drug Discovery*, vol. 23, no. 1, pp. 67–83, 2024. DOI: 10.1038/s41573-023-00827-x

[6] U. Şahin, K. Karikó, and Ö. Türeci, "mRNA-based therapeutics—developing a new class of drugs," *Nature Reviews Drug Discovery*, vol. 13, no. 10, pp. 759–780, 2014. DOI: 10.1038/nrd4278

[7] K. Leppek et al., "Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics," *Nature Communications*, vol. 13, no. 1, 2022, Art. no. 1536. DOI: 10.1038/s41467-022-28776-w

[8] D. M. Mauger et al., "mRNA structure regulates protein expression through changes in functional half-life," *Proceedings of the National Academy of Sciences*, vol. 116, no. 48, pp. 24 075–24 083, 2019. DOI: 10.1073/pnas.1908052116

[9] H. Zhang et al., "Algorithm for optimized mRNA design improves stability and immunogenicity," *Nature*, vol. 621, no. 7978, pp. 396–403, 2023. DOI: 10.1038/s41586-023-06127-z

[10] 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*, vol. 15, no. 3, pp. 1281–1295, 1987. DOI: 10.1093/nar/15.3.1281

[11] S. Patel et al., *mRNAutilus: Multi-objective-guided discrete generation of mRNA with optimized therapeutic properties*, arXiv preprint, 2026. DOI: 10.48550/arXiv.2605.31296 arXiv: 2605.31296v1 [q-bio.BM].

[12] H. Zhang et al., "Deep generative models design mRNA sequences with enhanced translational capacity and stability," *Science*, vol. 390, no. 6773, 2025, Art. no. eadr8470. DOI: 10.1126/science.adr8470

[13] S. Li, P. Chauvin, O. Gross, M. Bailey, and S. Jager, *mRNA-GPT: A generative model for full-length mRNA design and optimization*, bioRxiv preprint, 2026. DOI: 10.64898/2026.03.31.715707 biorxiv: 2026.03.31.715707.

[14] A. Medjmedj et al., "Evaluation of synthetic mRNA with selected UTR sequences and alternative poly(A) tail, in vitro and in vivo," *Molecular Therapy Nucleic Acids*, vol. 36, no. 3, 2025, Art. no. 102648. DOI: 10.1016/j.omtn.2025.102648

[15] U. Şahin and Ö. Türeci, "Personalized vaccines for cancer immunotherapy," *Science*, vol. 359, no. 6382, pp. 1355–1360, 2018. DOI: 10.1126/science.aar7112

[16] E. Blass and P. A. Ott, "Advances in the development of personalized neoantigen-based therapeutic cancer vaccines," *Nature Reviews Clinical Oncology*, vol. 18, no. 4, pp. 215–229, 2021. DOI: 10.1038/s41571-020-00460-2

2026 S.T. Yau High School Science Award 仅用于2026丘成桐中学科学奖论文公示

[17] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 27 730–27 744. DOI: 10.52202/068431-2011

[18] Z. Shao et al., *DeepSeekMath: Pushing the limits of mathematical reasoning in open language models*, arXiv preprint, 2024. DOI: 10.48550/arXiv.2402.03300 arXiv: 2402.03300 [cs.CL].

[19] N. Lambert et al., "Tülu 3: Pushing frontiers in open language model post-training," in *Second Conference on Language Modeling*, 2025. [Online]. Available: https://openreview.net/forum?id=i1uGbfHHpH

[20] D. Guo et al., "DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning," *Nature*, vol. 645, no. 8081, pp. 633–638, Sep. 17, 2025. DOI: 10.1038/s41586-025-09422-z

[21] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703

[22] N. A. O'Leary et al., "Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation," *Nucleic Acids Research*, vol. 44, no. D1, pp. D733–D745, 2016. DOI: 10.1093/nar/gkv1189

[23] S. E. Calvo, D. J. Pagliarini, and V. K. Mootha, "Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans," *Proceedings of the National Academy of Sciences*, vol. 106, no. 18, pp. 7507–7512, 2009. DOI: 10.1073/pnas.0810916106

[24] Y. Xiong, A. Wang, Y. Kang, C. Shen, C.-Y. Hsieh, and T. Hou, "mRNABERT: Advancing mRNA sequence design with a universal language model and comprehensive dataset," *Nature Communications*, vol. 16, no. 1, 2025, Art. no. 10371. DOI: 10.1038/s41467-025-65340-8

[25] L. Huang et al., "LinearFold: Linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search," *Bioinformatics*, vol. 35, no. 14, pp. i295–i304, 2019. DOI: 10.1093/bioinformatics/btz375

[26] C. Barreau, L. Paillard, and H. B. Osborne, "AU-rich elements and associated factors: Are there unifying principles?" *Nucleic Acids Research*, vol. 33, no. 22, pp. 7138–7150, 2005. DOI: 10.1093/nar/gki1012

[27] K. Karikó, M. Buckstein, H. Ni, and D. Weissman, "Suppression of RNA recognition by Toll-like receptors: The impact of nucleoside modification and the evolutionary origin of RNA," *Immunity*, vol. 23, no. 2, pp. 165–175, 2005. DOI: 10.1016/j.immuni.2005.06.008

[28] R. Verbeke, M. J. Hogan, K. Loré, and N. Pardi, "Innate immune mechanisms of mRNA vaccines," *Immunity*, vol. 55, no. 11, pp. 1993–2005, 2022. DOI: 10.1016/j.immuni.2022.10.014

[29] Y. Ichihara, Y. Jinnai, T. Morimura, M. Sakamoto, R. Mitsuhashi, and E. Uchibe, "MO-GRPO: Mitigating reward hacking of group relative policy optimization on multi-objective problems," *Transactions of the Association for Computational Linguistics*, vol. 14, pp. 1867–1882, 2026. DOI: 10.1162/tacl.a.782

[30] S.-Y. Liu et al., *GDPO: Group reward-decoupled normalization policy optimization for multi-reward RL optimization*, arXiv preprint, 2026. DOI: 10.48550/arXiv.2601.05242 arXiv: 2601.05242 [cs.CL].

[31] Q. Yu et al., "DAPO: An open-source LLM reinforcement learning system at scale," in *Advances in Neural Information Processing Systems*, vol. 38, Curran Associates, Inc., 2025, pp. 113 222–113 244. DOI: 10.52202/085713-3775

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖公示

[32] DeepSeek-AI et al., *DeepSeek-V3.2: Pushing the frontier of open large language models*, arXiv preprint, 2025. DOI: 10.48550/arXiv.2512.02556 arXiv: 2512.02556 [cs.CL].

[33] V. Agarwal and D. R. Kelley, "The genetic and biochemical determinants of mRNA degradation rates in mammals," *Genome Biology*, vol. 23, no. 1, 2022, Art. no. 245. DOI: 10.1186/s13059-022-02811-x

[34] J. Liu et al., *Muon is scalable for LLM training*, arXiv preprint, 2025. DOI: 10.48550/arXiv.2502.16982 arXiv: 2502.16982 [cs.LG].

[35] K. Jordan et al. "Muon: An optimizer for hidden layers in neural networks," Accessed: Sep. 15, 2026. [Online]. Available: https://kellerjordan.github.io/posts/muon/

[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[37] D. Zheng et al., "Predicting the translation efficiency of messenger RNA in mammalian cells," *Nature Biotechnology*, vol. 44, no. 5, pp. 783–796, 2026. DOI: 10.1038/s41587-025-02712-x

[38] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, 2019, pp. 2623–2631. DOI: 10.1145/3292500.3330701

[39] D. G. Gibson, L. Young, R.-Y. Chuang, J. C. Venter, C. A. Hutchison III, and H. O. Smith, "Enzymatic assembly of DNA molecules up to several hundred kilobases," *Nature Methods*, vol. 6, no. 5, pp. 343–345, 2009. DOI: 10.1038/nmeth.1318

[40] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979. [Online]. Available: https://www.jstor.org/stable/4615733

[41] L. McInnes, J. Healy, and J. Melville, *UMAP: Uniform manifold approximation and projection for dimension reduction*, arXiv preprint, 2018. DOI: 10.48550/arXiv.1802.03426 arXiv: 1802.03426 [stat.ML].

[42] M. Kozak, "An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs," *Nucleic Acids Research*, vol. 15, no. 20, pp. 8125–8148, 1987. DOI: 10.1093/nar/15.20.8125

[43] M. Kozak, "Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes," *Cell*, vol. 44, no. 2, pp. 283–292, 1986. DOI: 10.1016/0092-8674(86)90762-2

[44] G. Kudla, L. Lipinski, F. Caffin, A. Helwak, and M. Zylicz, "High guanine and cytosine content increases mRNA levels in mammalian cells," *PLoS Biology*, vol. 4, no. 6, 2006, Art. no. e180. DOI: 10.1371/journal.pbio.0040180

[45] T. Tuller and H. Zur, "Multiple roles of the coding sequence 5' end in gene expression regulation," *Nucleic Acids Research*, vol. 43, no. 1, pp. 13–28, 2015. DOI: 10.1093/nar/gku1313

[46] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990. DOI: 10.1016/S0022-2836(05)80360-2

2026 S. T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

Appendix A Experimental Setup

A.1 Model Training

Supervised training details. The BASE model is trained with the scalable Muon variant proposed in Moonlight [34], which augments the original Muon optimizer [35] with weight decay and parameter-wise update-scale adjustment. We use learning rate 10^{-3}, weight decay 0.1, equivalent batch size 1,024, and linear warmup over 1,000 steps followed by cosine annealing to 1% of the peak learning rate. We train for 20 epochs and use the final checkpoint as BASE; SFT is initialized from the BASE checkpoint and trained on 37,369 NM-prefixed RefSeq mRNAs without upstream open reading frames (uORFs), with batch size 128 and peak learning rate 5×10^{-5} for 3 epochs.

RL post-training details. ProMORNA-RL is initialized from the SFT checkpoint, and the reference policy π_{ref} is fixed as a frozen copy of the SFT model. Then we employ our stabilized MO-GRPO post-training procedure (Section 2.3.3) as a general protein-conditioned optimization stage: we randomly sample 4,000 protein prompts from the training set (firefly luciferase target excluded), pulling a batch of 4 protein prompts from this pool at each RL step. For each prompt, the policy generates a group of $G = 32$ complete mRNA rollouts for policy updates. We train for 1,000 RL steps using AdamW [36] with constant learning rate 10^{-5} and linear warmup over 50 steps. The KL penalty coefficient is set to $\beta = 0.02$ and the advantage clipping bound is set to $c = 3.0$. The policy ratio clipping term is naturally omitted since we only perform one optimizer step per batch. The objective weights are set to $(w_{\text{stability}}, w_{\text{TE}}, w_{\text{MFE}}, w_{\text{U}}, w_{\text{UTR}}) = (1.0, 1.0, -0.5, -0.5, 0.5)$, where negative weights indicate minimization objectives. The damping factors are set to $(\delta_{\text{stability}}, \delta_{\text{TE}}, \delta_{\text{MFE}}, \delta_{\text{U}}, \delta_{\text{UTR}}) = (0.2, 0.2, 0.01, 0.01, 0.01)$. Invalid sequences receive the maximal penalty before post-centralization. The stability and translation efficiency predictors are frozen throughout RL training.

Reward predictor training. Predicted stability and predicted translation efficiency serve as the two primary functional objectives in our reward design. Since these quantities cannot be directly computed from sequence alone, we train two supervised proxy predictors to estimate them from mRNA sequences. Each mRNA sequence is first encoded by a frozen mRNABERT model [24], and the resulting token-level representations are aggregated into a 768-dimensional sequence representation by global average pooling. This reduces the two prediction tasks to independent mappings from mRNABERT sequence embeddings to scalar property values. The translation-efficiency and stability predictors are then trained with ridge regression on human measurements from the transcriptome-wide atlas [37] and the mammalian mRNA degradation compendium [33], respectively.

The ridge regularization strength α is optimized separately for each task using Optuna [38]. For each predictor, we run 100 hyperparameter trials and sample α from the logarithmic range $[10^{-2}, 10^{5}]$. The hyperparameter search primarily maximizes validation Spearman's rank correlation coefficient (SRCC). When two trials have validation SRCC values differing by no more than 10^{-4}, we select the one with the lower validation mean absolute error (MAE). To reduce variance in hyperparameter selection, we use repeated 5-fold cross-validation with three repeats.

Prediction task	SRCC	MAE	α
Stability	0.635	0.608	134.27
Translation efficiency	0.744	0.364	82.80

Table 1: Cross-validation performance of the mRNABERT-based proxy predictors.

2026 Shanghai High School Science Award
仅用于2026上海高中科学奖论文公示

A.2 *In Silico* Experimental Setup

Target protein. For a case study, we evaluate all methods on firefly luciferase (Fluc2P), following the reporter choice in the GEMORNA study [12]. Only the amino-acid sequence of the target protein is provided to ProMORNA at inference time, and no wild-type mRNA template is used. To avoid potential data leakage, transcripts corresponding to the target protein, if present, are removed from all training datasets and the RL post-training prompt pool. It is worth noting that the framework is protein-agnostic and can generate full-length mRNAs for any given protein sequence.

Baselines. We compare ProMORNA-RL with three baselines. BASE denotes the supervised encoder–decoder model trained on the full RefSeq training set. SFT denotes the BASE model further fine-tuned on the filtered `NM_` subset. The GEMORNA baseline is constructed by combining the 12 GEMORNA-designed 5' UTRs, the four GEMORNA firefly-luciferase CDSs, and the 10 GEMORNA-designed 3' UTRs from their released supplementary data [12]. Their Cartesian product yields $12 \times 4 \times 10 = 480$ full-length GEMORNA mRNA candidates. Notably, the GEMORNA-derived sequences used for comparison are selected through wet-lab screening in their original experimental pipeline, and thus represent a carefully selected, high-performing set. For a fair comparison, all candidates from different models are evaluated using the same validity checker and the same evaluation metrics.

Candidate generation. For the three ProMORNA variants, sequences are generated using constrained stochastic decoding with temperature $\tau = 1.0$, top-p filtering disabled, and a maximum decoder length of 1,024 tokens. Each sequence is forced to begin with `AGG` to ensure fair comparison with other methods. We repeat this generation procedure until 1,000 valid candidates are produced. A generated sequence is accepted only if it passes these basic biological filters:

- $|5'\text{ UTR}| \geq 20$ and $|3'\text{ UTR}| \geq 30$;
- the CDS correctly translates to the target protein;
- no uORF is present, i.e. `"AUG"` $\notin$ 5' UTR.

For GEMORNA, 480 candidates are constructed from the Cartesian product of its released 5' UTR, CDS, and 3' UTR elements.

Evaluation protocol. All reported metric distributions are computed on valid and unique candidates. For each method, we evaluate (i) reward-related metrics, including predicted stability, predicted translation efficiency, normalized MFE, U-content, 5' UTR length, and 3' UTR length; (ii) diagnostic sequence-property metrics, including GC content, Codon Adaptation Index (CAI), the normalized MFE of the 5' leader region, the normalized MFE of the remaining mRNA body, external GEMORNA UTR scores, and 5' and 3' UTR Max Identity. For RL learning curves, we report the average metric value over all valid generated candidates in every training step.

A.3 *In Vitro* Validation

Plasmid Construction. The plasmids used in this study are constructed via restriction enzyme cloning and Gibson assembly [39].

Reaction Assembly. In vitro transcription (IVT) reactions are assembled as described in Table 2. After brief centrifugation, the reaction is incubated at 37°C for 1.5–2 hours.

DNA Digestion. DNase I (10 U) is added to each reaction, followed by incubation at 37°C for 30 minutes.

2026 S.T.Yau High School Science Award
仅用于2026丘成桐中学科学奖公示

Component	Amount
RNase-free H_2O	Up to 20 μL
10× Transcription Buffer	2 μL
ATP / CTP / GTP / N1-Me-pUTP (100 mM each)	2 μL each
CAP GAG 3011 (100 mM)	2 μL
Template DNA	1 μg
T7 RNA Polymerase Mix	2 μL
Total Reaction Volume	**20 μL**

Table 2: Composition of the *in vitro* transcription reaction.

RNA Purification. The RNA product is purified via lithium chloride precipitation. An equal volume of lithium chloride precipitation solution (5 M) is added to the reaction product and mixed thoroughly. The mixture is precipitated at −20°C for 30 minutes and then centrifuged at 12,000 RPM at 4°C for 15 minutes. The supernatant is discarded. The RNA pellet is washed with 200 μL of 70% ice-cold ethanol and centrifuged at 12,000 RPM at 4°C for 15 minutes. The supernatant is discarded and the washing step is repeated 2–3 times. The pellet is air-dried with the lid open for 5–10 minutes. Once completely dry, the RNA pellet is dissolved in 100–200 μL of RNase-free ddH_2O or an appropriate buffer. The RNA concentration is measured, aliquoted, and stored at −80°C.

mRNA Transfection. HEK293T cells are cultured in DMEM (Gibco, 6126249) supplemented with 10% FBS (Vazyme, 7E951A5), 100 units/mL penicillin, and 100 μg/mL streptomycin. One day prior to transfection, cells are seeded into 96-well plates at a density of 20,000 cells per well. Transfection is performed using Lipofectamine® RNAiMAX Reagent (Invitrogen, 13778-030) following the manufacturer's protocol.

Cell Lysis. To measure firefly luciferase activity, 100 μL of reporter cell lysis buffer is added to each well to ensure complete cell lysis. After lysis, the lysate is centrifuged at 10,000–15,000×g for 3–5 minutes, and the supernatant is collected for subsequent measurement.

Detection Reagent Preparation. Firefly luciferase detection reagent is thawed and equilibrated to room temperature. The chemiluminescence reader (or a multifunctional microplate reader with chemiluminescence detection capability) is operated according to the manufacturer's instructions, with appropriate interval and measurement times set based on experimental requirements.

Luminescence Measurement. For each sample, 80 μL of supernatant is transferred to the assay plate. An equal volume of reporter cell lysis buffer serves as a blank control. Subsequently, 100 μL of firefly luciferase detection reagent is added to each well, mixed thoroughly by pipetting, and the luminescence signal is immediately measured.

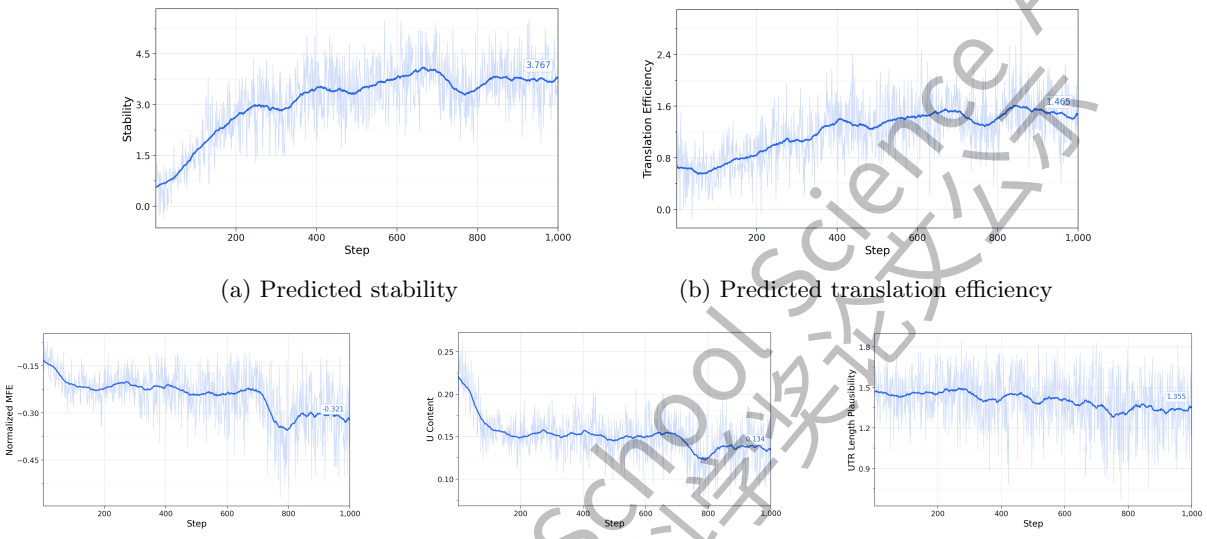
Data Processing and Statistical Analysis. We average first-cycle readings from technical wells within each construct, experiment, and time point (four wells in experiment 1 and three in experiments 2 and 3). In this retrospective analysis, the principal activity measure is the trapezoidal AUC of the raw counts/s curve over 12–72 h, calculated before normalization. All five time points are retained, with no additional background subtraction. Candidate/control AUC ratios are paired within experiment. We report geometric means and pointwise 95% confidence intervals obtained from the three log ratios using the t distribution ($n = 3, df = 2$). These intervals assume approximately normal log ratios and are not adjusted for multiplicity or candidate selection. Two-sided t tests assess whether the mean log ratio differs from zero. Holm correction [40] controls family-wise error separately within the seven

2026 S.T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

candidate–Benchmark comparisons (B7; screening) and the seven candidate–GMR-FL-F5 comparisons (G7; exploratory). Time courses are descriptive.

Appendix B Reinforcement Learning Dynamics

In this section, we investigate how RL post-training alters the generator. Figure 10 presents the learning curves for the five reward-related metrics over 1,000 RL steps, averaged across valid rollouts from protein prompts sampled during general RL post-training. The light traces represent step-level variance, while the bold curves depict the smoothed trends.



(a) Predicted stability

(b) Predicted translation efficiency

(c) Normalized MFE

(d) U-content

(e) UTR length plausibility

Figure 10: The learning curves of RL post-training on five reward metrics.

RL post-training successfully optimizes the two primary functional proxy objectives. Notably, predicted stability increases rapidly and steadily throughout most of the training. Predicted translation efficiency experiences an initial decline due to early trade-offs but subsequently recovers, rising steadily before gradually approaching a plateau. Concurrently, normalized MFE decreases, indicating the formation of more stable full-transcript secondary structures, and U-content drops sharply, aligning consistently with the reward design. The UTR length plausibility score slightly lowered, which matches our expectations for it as a regularization term. Overall, these trajectories demonstrate that multi-objective RL post-training can effectively improve multiple rewards simultaneously, despite the noise introduced by stochastic full-length rollouts.

To understand the evolution of sequence composition during training, Fig. 11 visualizes the generated mRNAs using a UMAP projection [41] of 1,024-dimensional 5-mer frequency features. The projection reveals a striking bifurcation in the latent space: early RL samples cluster tightly around the original SFT baseline on the right, whereas progressing training steps drive the candidates to migrate into a completely separate, vastly expanded region on the left. This distinct spatial shift underscores that RL post-training does not simply exploit the supervised model's local vicinity, but actively and extensively explores novel sequence-composition paradigms. While this compositional shift does not by itself establish biological quality, pairing this structural migration with the previously discussed learning curves demonstrates that on-policy optimization effectively navigates away from the SFT distribution to uncover highly distinct sequences with superior reward profiles.

The Kozak sequence is a well-established cis-regulatory element in eukaryotic mRNA that strongly influences translation initiation efficiency [42, 43]. According to the canonical Kozak consensus sequence,

2026 S. T. Yau High School Science Award
应用于2026年丘成桐中学科学奖公示

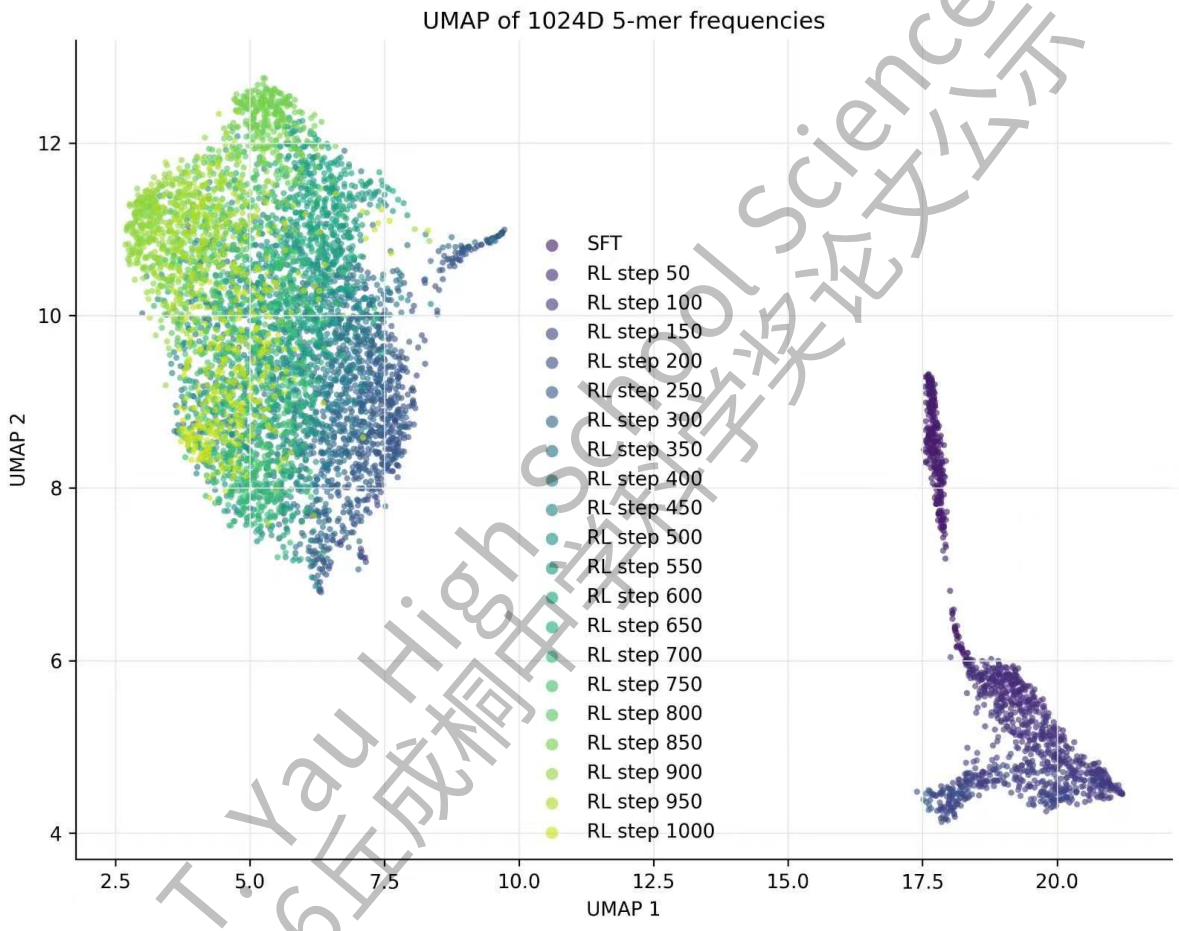


Figure 11: UMAP visualization of 5-mer features from mRNAs generated during RL post-training.

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

the nucleotide at position −3 relative to the start codon should preferably be a purine, with adenine (A) being the most optimal, while the −1 position is ideally a cytosine (C). We investigated whether our RL-optimized models autonomously learned to respect these biological constraints, despite receiving no explicit Kozak-related inductive bias during training. The results are illustrated in Fig. 12.

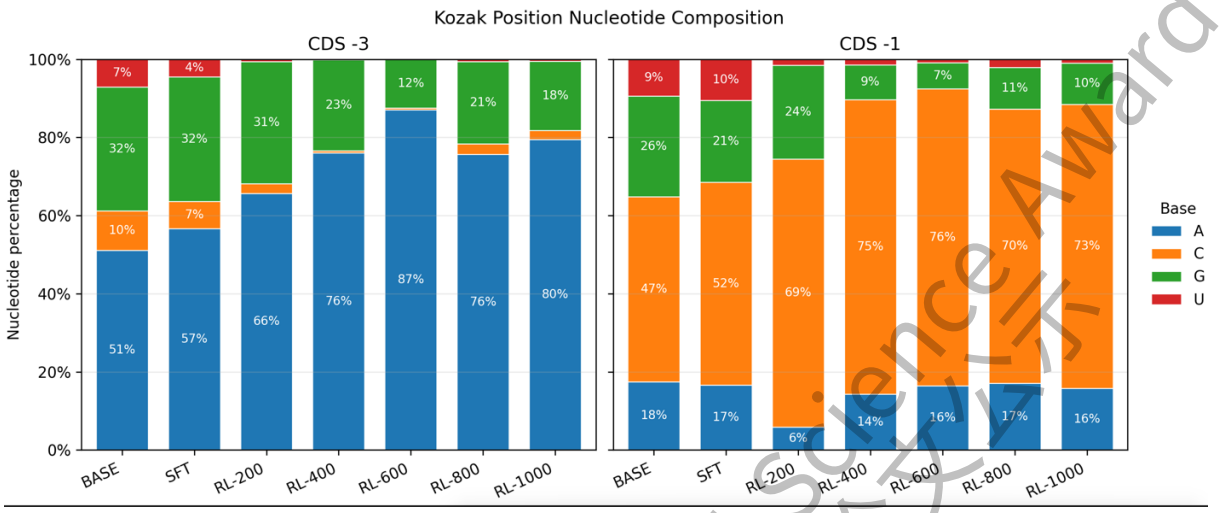


Figure 12: RL shifts nucleotide compositions at Kozak positions.

At the -3 position, the model exhibits a clear progressive enrichment for purines (A/G) across training stages, with a striking shift toward the optimal adenine. The BASE model demonstrates a moderate baseline preference (A+G = 83%), which is further reinforced by supervised fine-tuning (SFT: 89%). Strikingly, RL training not only drives the combined purine content near 97–99% (from RL-400 onward), but it also autonomously discovers the superiority of adenine: the frequency of `A` alone surges from 57% in SFT to a peak of 87% at RL-600, while `G` concurrently drops from 32% to 12%.

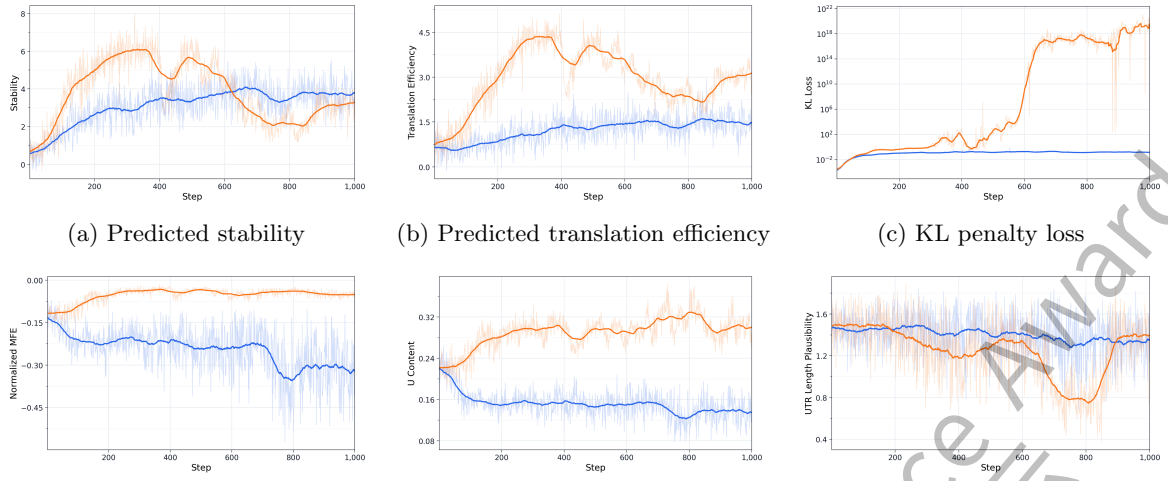
Similarly, at the -1 position, the trend toward Kozak compliance is highly pronounced. The BASE model places a cytosine at this position only 47% of the time, and SFT improves this to 52%. However, RL training yields a dramatic shift: C content jumps to 69% at RL-200, peaks at 76% at RL-600, and remains robustly high across all subsequent checkpoints. Together, these results provide evidence of biological interpretability within our RL pipeline.

Appendix C MO-GRPO Ablation

To evaluate the necessity of metric-level advantage standardization and our stabilization mechanisms, we replace our MO-GRPO formulation in RL post-training with a scalarized GRPO-style ablation, in which the reward is calculated by the weighted sum of the five metrics defined in Section 2.3.2, using exactly the same metric weights reported in Appendix A.1, and the advantage is given by standardized scalar rewards. Invalid sequences are assigned a reward of zero. Other hyperparameters remain unchanged. The comparison of the two learning curves is shown in Fig. 13.

Consistent with recent theoretical analyses of variance dominance and reward signal collapse in multi-objective GRPO [29, 30], because metric weights aren't divided by their group standard deviations in the scalarized GRPO, predicted stability and translation efficiency dominated the reward calculation at early stages of RL, resulting in severe reward hacking (Figs. 13a and 13b). This caused the policy to diverge largely from the reference SFT policy, and finally training collapsed because of the exploding KL divergence penalty (Fig. 13c). Due to reward hacking, only the two mRNABERT predictions improved, while all other metrics deteriorated (Figs. 13d to 13f). On the other hand, MO-GRPO training was very stable, benefiting from metric-level advantage calculation [29, 30] combined with our dynamic damping,

2026 S.T. Yau High School Science Award
项目于2026日成稿学术论文公示



(a) Predicted stability (b) Predicted translation efficiency (c) KL penalty loss

(d) Normalized MFE (e) U-content (f) UTR length plausibility

Figure 13: MO-GRPO versus scalarized GRPO on ProMORNA post-training. The orange curves are from scalarized GRPO, while the blue curves are the RL post-training curves reported in Appendix B.

extreme-value clipping, and validity-masked penalty methods.

Appendix D Additional mRNA Property Analysis

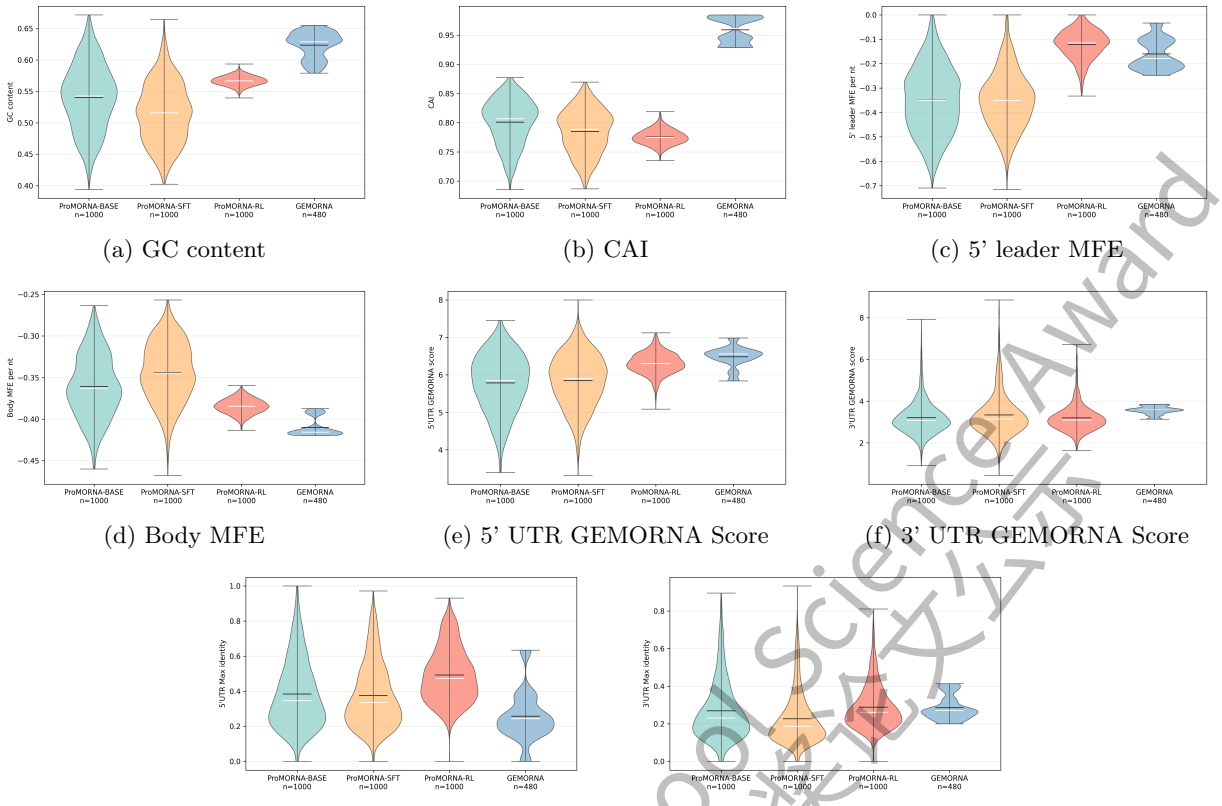
To better understand the sequence-level mechanics driving these results, Fig. 14 details additional sequence properties. Although not primary reward objectives, these metrics provide valuable mechanistic insights into how ProMORNA shapes nucleotide and codon usage and alters regional mRNA secondary structure.

As illustrated in Fig. 14a, ProMORNA-RL produces a moderate increase in median GC content relative to both the BASE and SFT models, accompanied by a dramatic reduction in variance. While higher GC content is linked to increased mRNA abundance in mammalian cells [44], excessive GC enrichment can deleteriously impact local secondary structure, perturb optimal codon context, and hinder downstream sequence manufacturability. ProMORNA's moderate increase aligns with a balanced optimization strategy, avoiding the extremes of a single-proxy focus.

A similar multi-objective trade-off is observed in the Codon Adaptation Index (CAI) [10] (Fig. 14b). While GEMORNA exhibits near-maximal CAI values due to its strong preference for high-frequency synonymous codons, ProMORNA-RL converges to a tightly concentrated, moderate CAI regime around 0.78, which is slightly lower than those of the BASE and SFT models. Because CAI is not explicitly included in our reward formulation, this distribution indicates that RL post-training actively trades off greedy codon optimality to satisfy competing global constraints, such as minimizing U-content, enhancing full-length thermodynamic stability, and maintaining favorable local folding topologies [5, 8]. Notably, despite having a lower CAI than GEMORNA, ProMORNA-RL achieves superior predicted translation rates under the independent Saluki evaluator and strong *in vitro* reporter activity, demonstrating that maximizing heuristic codon usage alone is neither necessary nor sufficient for optimal protein expression.

Figures 14c and 14d map region-specific mRNA secondary structure by assessing normalized MFE of different regions. Biologically, a higher MFE in the 5' leader region (the 5' UTR and first 10 CDS nucleotides) indicates looser secondary structure, which is known to facilitate efficient ribosome scanning [8, 45]. Conversely, tighter secondary structures in the remainder of the mRNA body (the rest of the CDS and the 3' UTR) broadly enhance transcript stability and protein output, corresponding to a lower MFE [7]. Compared to the BASE and SFT models, ProMORNA-RL closely aligns with these biological design principles: it shifts the 5' leader MFE upward to minimize translation initiation barriers, while simultaneously driving the body MFE downward to stabilize the rest of the transcript. Although

2026 S. T. Yau High School Science Award 仅用于2026丘成桐中学科学奖论文公示



(a) GC content (b) CAI (c) 5' leader MFE

(d) Body MFE (e) 5' UTR GEMORNA Score (f) 3' UTR GEMORNA Score

(g) 5' UTR Max Identity (h) 3' UTR Max Identity

Figure 14: Additional sequence-property analysis of valid, unique firefly-luciferase mRNA candidates. The GEMORNA UTR scores are computed using the external UTR scoring models from GEMORNA [12].

GEMORNA achieves more extreme MFE values in both regions, likely due to its reliance on a heavily curated sequence pool, ProMORNA autonomously learns to optimize these regional folding topologies while retaining structural diversity.

Figures 14e and 14f evaluate the generated sequences using external GEMORNA UTR scoring models, which estimate UTR quality based on predictors trained for high Mean Ribosome Load in 5' UTRs and high stability in 3' UTRs [12]. While GEMORNA exhibits concentrated, high-scoring distributions due to sampling UTRs from its own pre-optimized pool, ProMORNA-RL achieves a highly competitive 5' UTR score distribution. Notably, it demonstrates significant improvement and reduced variance relative to the BASE and SFT models, all while maintaining broader sequence diversity.

We also assessed the sequence similarities between generated and natural UTRs using BLAST [46]. A low max identity score (MIS) indicates minimal homology between the AI-generated UTRs and naturally occurring UTRs in the reference database. As shown in Figs. 14g and 14h, the MIS distributions for the 5' and 3' UTRs reveal that the ProMORNA models generate sequences with low to moderate homology to known natural transcripts. This confirms that the pipeline performs genuine *de novo* generation rather than merely memorizing sequences from reference databases.

Furthermore, ProMORNA-RL's two regulatory regions display distinct sequence-space dynamics. For the 5' UTR (Fig. 14g), ProMORNA-RL demonstrates a moderate upward shift in median MIS compared to the BASE and SFT models; for the 3' UTR (Fig. 14h), it maintains a persistently low sequence similarity. Taken together, these regional differences may highlight a sophisticated, region-specific optimization strategy within ProMORNA-RL: it autonomously converges toward natural-like motifs where translational initiation constraints are strict (5' UTR), while broadly exploring entirely novel, artificial sequence spaces where structural flexibility permits (3' UTR).

2026 c. T. Yau High School Science Award 仅用于2026江苏省中学科学奖论文公示