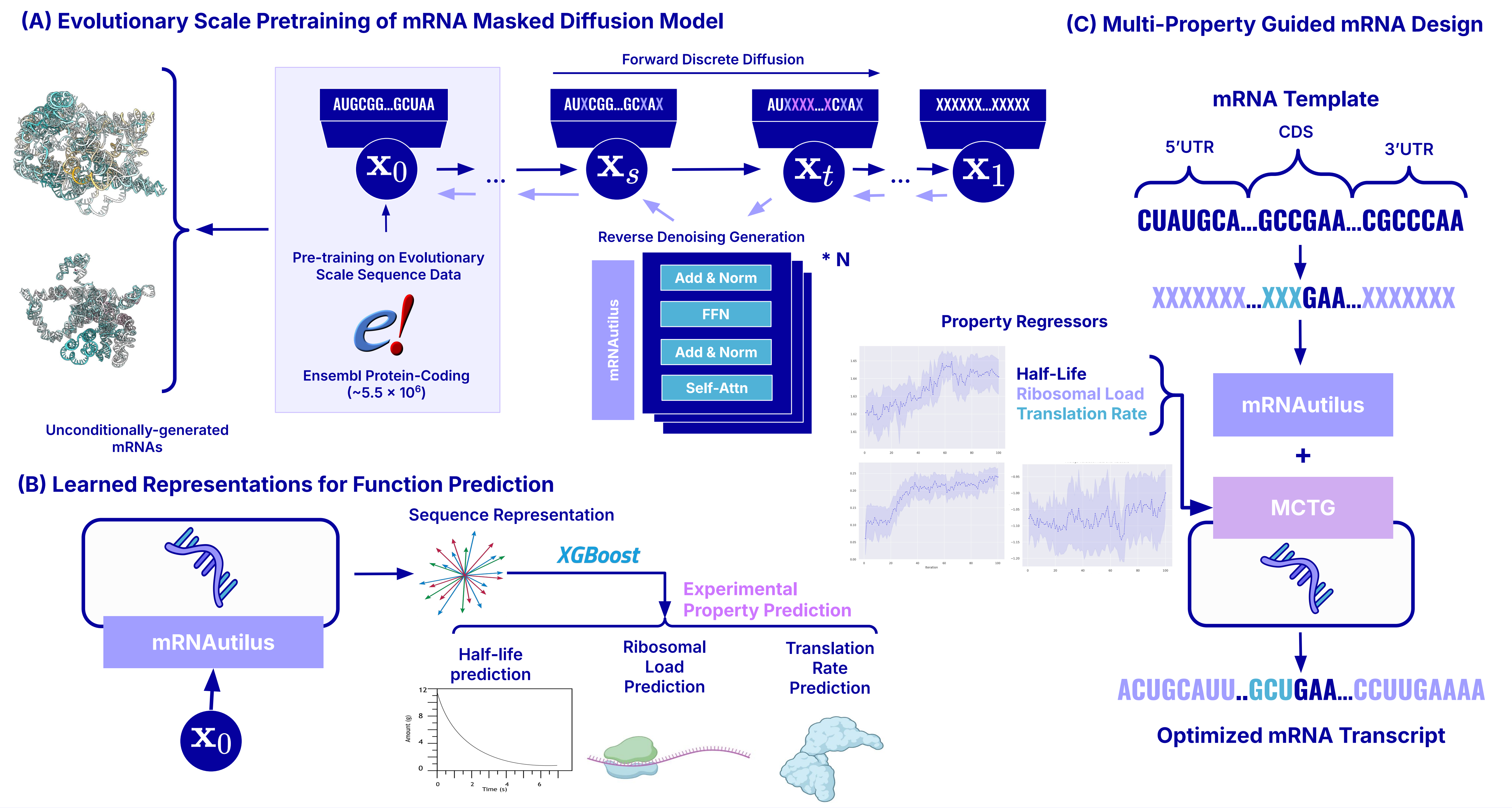
Multi-Objective-Guided Generative Design of mRNA with Therapeutic Properties

Sawan Patel[1], **Sophia Tang**[2], Yinuo Zhang[3], Pranam Chatterjee[2,4 †], Sherwood Yao[1 †]

[1]Atom Bioworks
[2]Department of Computer and Information Science, University of Pennsylvania
[3]Center of Computational Biology, Duke-NUS Medical School
[4]Department of Bioengineering, University of Pennsylvania
[†]Correspondence to: s.yao@atombioworks.com | pranam@seas.upenn.edu

Motivation

1. **Therapeutic mRNAs require co-optimality across several properties of interest.** Designing mRNAs for delivery, such as in mRNA-LNP delivery platforms, is dependent on the mRNA's potential to translate into the antigen of interest and trigger an immune response. For this, properties such as half-life, polysome formation propensity, immunogenicity, translation rate, and more must be considered when evaluating the efficacy of an mRNA payload.
2. **UTR design and codon optimization should be done simultaneously.** Typically, an existing set of 5'/3' untranslated regions (UTRs) is paired with an mRNA open reading frame (ORF) with assumed compatibility. However, interactions between the ORF and UTRs can limit mRNA efficacy. Novel UTRs should be designed in the context of an ORF.
3. **Typical mRNA engineering does not consider the diversity of the mRNA sequence space.** Canonical codon optimization consists of referencing species-specific codon usage tables for replacing rare codons. Though sensible, avoiding rare codons is not always beneficial. Moreover, stitching existing UTRs to novel open reading frames leaves the UTR design space unexplored. Further combinatorics, whether during codon optimization, UTR design, or both, ought to be explored in a responsible design to avoid bias.



(A) Evolutionary Scale Pretraining of mRNA Masked Diffusion Model

Unconditionally-generated mRNAs

Pre-training on Evolutionary Scale Sequence Data

Ensembl Protein-Coding (~5.5 × 10⁶)

Forward Discrete Diffusion

Reverse Denoising Generation

AUGCGG...GCUAA — $\mathbf{x}_0$ → ... → AUXCGG...GCXAX — $\mathbf{x}_s$ → AUXXXX...XCXAX — $\mathbf{x}_t$ → ... → XXXXXX...XXXXX — $\mathbf{x}_1$

mRNAutilus: Add & Norm, FFN, Add & Norm, Self-Attn, * N

(B) Learned Representations for Function Prediction

mRNAutilus — $\mathbf{x}_0$

Sequence Representation

XGBoost

Experimental Property Prediction

Half-life prediction

Ribosomal Load Prediction

Translation Rate Prediction

(C) Multi-Property Guided mRNA Design

mRNA Template

5'UTR — CDS — 3'UTR

CUAUGCA...GCCGAA...CGCCCAA

XXXXXXX...XXXGAA...XXXXXXX

mRNAutilus + MCTG

Property Regressors: Half-Life, Ribosomal Load, Translation Rate

ACUGCAUU..GCUGAA...CCUUGAAAA

Optimized mRNA Transcript

Masked Diffusion Model

We model our sequence generation problem as a masked discrete diffusion process **(1)**. We sample from the parameterized reverse posterior of the diffusion process **(2)**, where a neural network is trained to reverse the noising process **(3)**.

$$p^{\text{mask}}_{t|0}(x_t|x_0) = \text{Cat}(x_t; tx_0 + (1-t)M) \tag{1}$$

$$p^{\theta}_{s|t}(x_s|x_t) = \begin{cases} (1-\frac{s}{t})x_0^{\theta} + \frac{s}{t}M & x_t = M \\ x_s & x_s \neq M \end{cases} \tag{2}$$

$$\mathcal{L}^{\infty}_{\text{NELBO}} = \mathbb{E}_{t, p_{t|0}(x_t|x_0), p_0(x_0)}\left[-\frac{1}{t}\log\langle x_0, x_0^{\theta}\rangle\right] \tag{3}$$

Monte-Carlo Tree Guidance

Selection: Start from a fully masked sequence (root node) and follow a sequence of optimal unmasking steps to a leaf node (unexpanded partially masked sequence)

Expansion: From the probability distribution generated from the trained diffusion model, apply Gumbel noise and sample M distinct partially unmasked sequences.

Rollout: From each partially unmasked child sequence, use greedy unmasking to fully unmask the sequence for scoring. Feed the unmasked sequences into a set of K classifiers to determine Pareto optimality.

Backpropagation: Calculate a reward vector of the fraction of the Pareto-optimal set that a sequence has a greater or equal score. Add the rewards across child nodes and add to the rewards of all predecessor nodes

Unconditional Generation

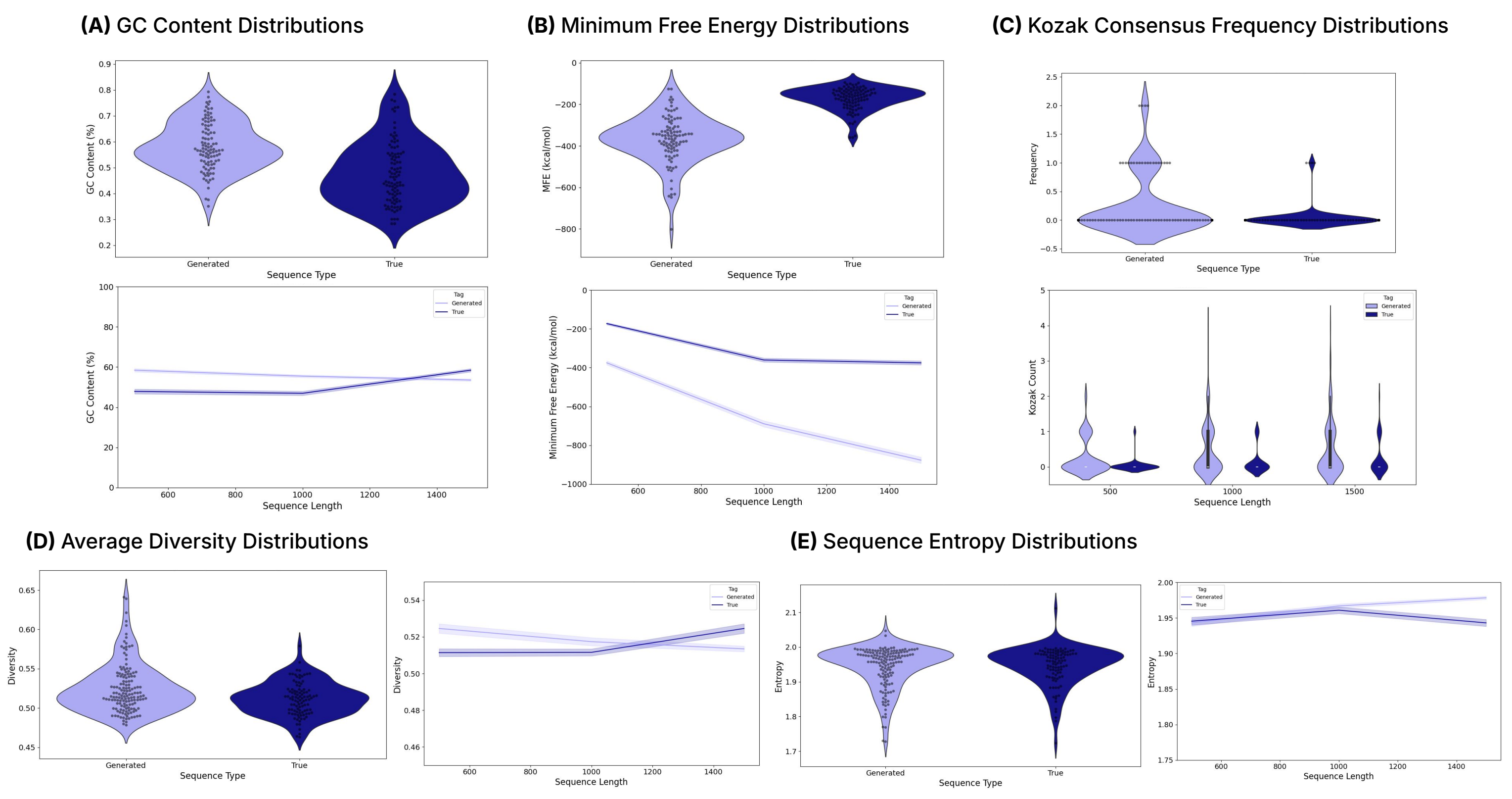


Fig 1. Unconditional, unguided generation of mRNA. Distributions for sequences of length 500 are shown as violin plots, for both generated (**violet**) and to natural mRNAs (**navy**). Plots evaluate sequence **(A)** GC content distributions (%), **(B)** predicted minimum free energy (kcal/mol), **(C)** kozak consensus sequence frequency, **(D)** average pairwise sequence diversity, and **(E)** sequence entropy (bits). Line plots below each panel show averages across across sequence lengths.

Latent Representation Analysis

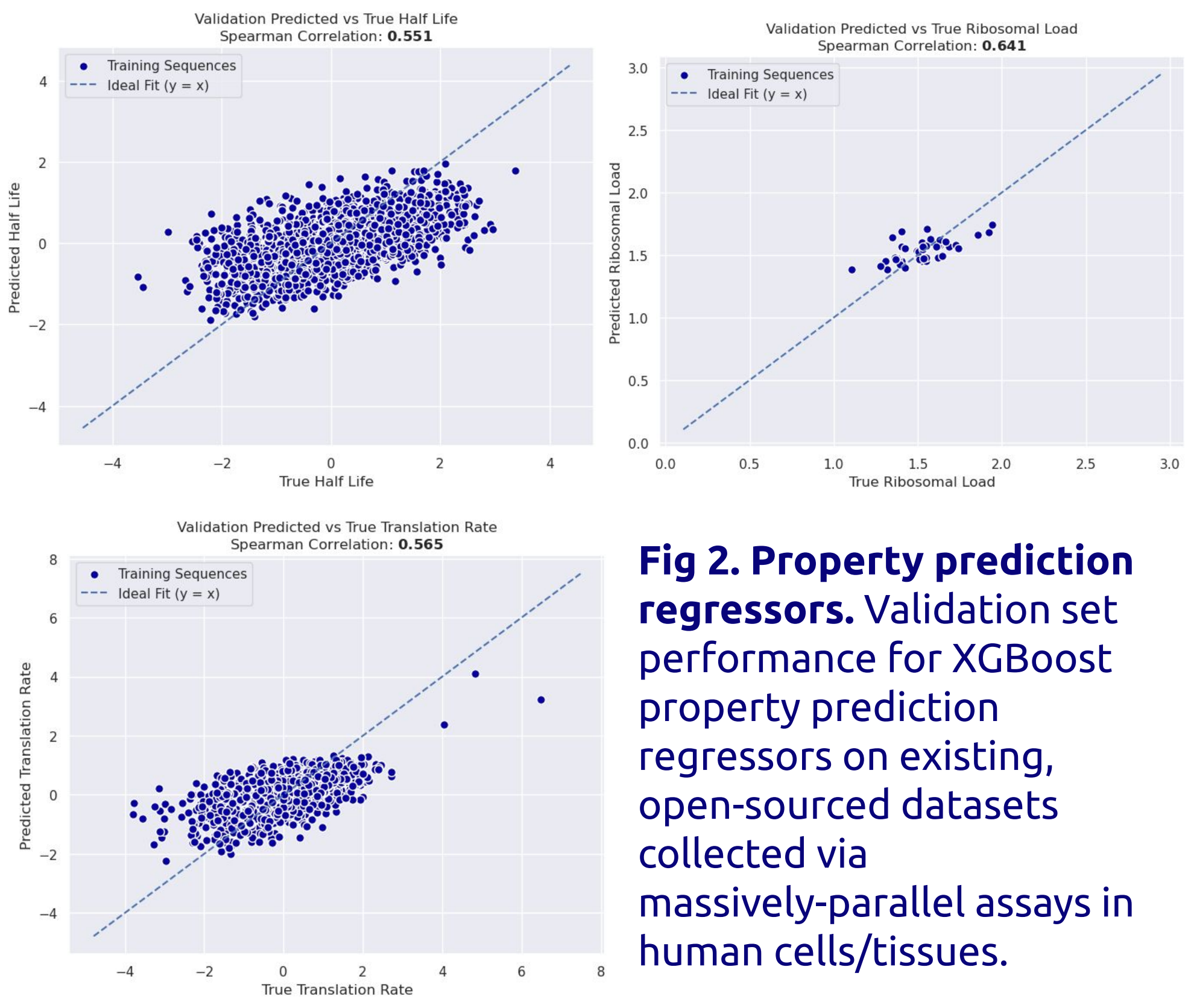
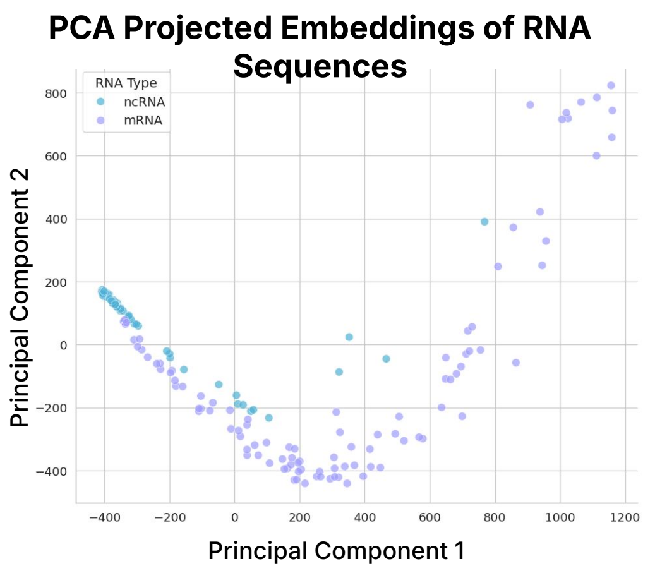


Fig 2. Property prediction regressors. Validation set performance for XGBoost property prediction regressors on existing, open-sourced datasets collected via massively-parallel assays in human cells/tissues.

Fig 3. Principal component analysis of mRNA and ncRNA embeddings using mRNAutilus. 100 natural mRNAs and ncRNAs were embedded and projected onto the first 2 PCs of the concatenated embedding matrix. mRNAs (**violet**) and ncRNA (**blue**) clusters are separable in the latent space.



Multi-Objective-Guided Generation

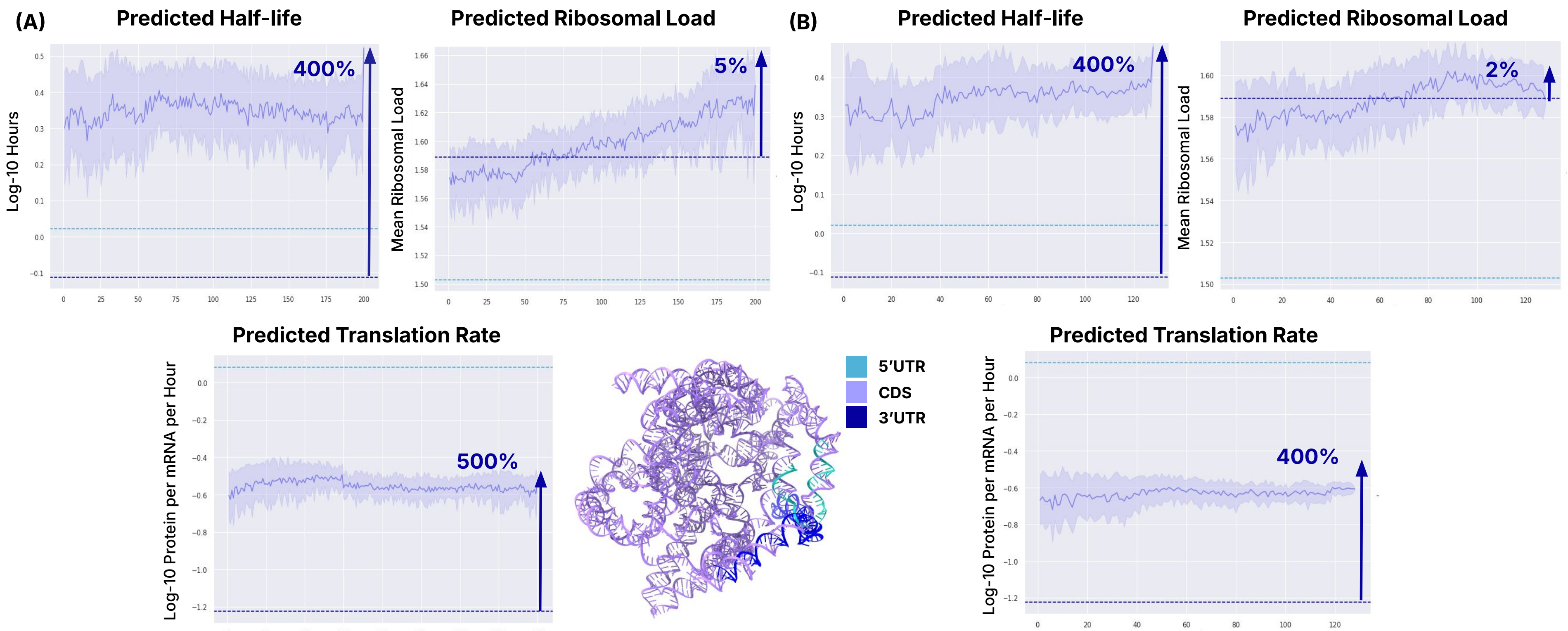
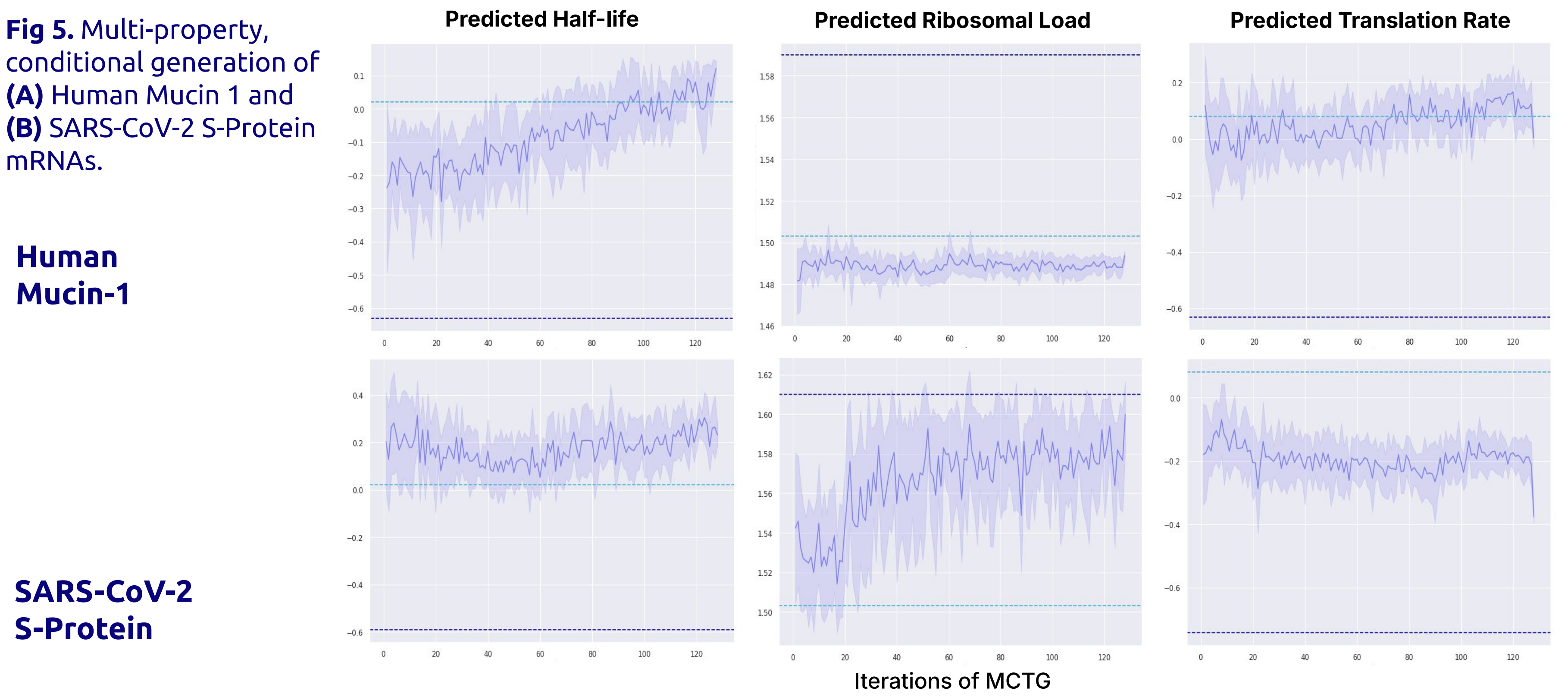


Fig 4. Multi-property, conditional generation of *P. pyralis* luciferase mRNA. A set of *Fluc* mRNAs is generated and optimized across properties of interest. Codon optimization and UTR design are done in parallel over Monte Carlo Tree Guidance time course. Codon optimization is done **(A)** up to the 50th codon and **(B)** for the entire CDS. Wild-type mRNA (**navy**) and classifier median scores (**teal**) are shown as horizontal lines.

Fig 5. Multi-property, conditional generation of **(A)** Human Mucin 1 and **(B)** SARS-CoV-2 S-Protein mRNAs.



Template Gene	Half-Life (log-10 hours; ↑)		Ribosome Profiling (MRL) (↑)		Translation Rate (log-10 scale; ↑)	
	wild-type	designed	wild-type	designed	wild-type	designed
Fluc	−0.112	**0.537**	1.58	**1.62**	−1.22	**−0.451**
SARS-CoV-2-S-Protein	−0.590	**0.297**	**1.61**	**1.61**	−0.741	**−0.0562**
MUC1	−0.092	**0.026**	**1.59**	1.49	−0.630	**−0.0389**

Table 1. Generated mRNA property scores in comparison to WT mRNAs for each gene.

Conclusions

1. We introduce **mRNAutilus**, a masked diffusion model for generation of **diverse, naturalistic, thermodynamically stable** mRNA sequences *de novo*.
2. We utilize the rich, learned latent space of **mRNAutilus** to predict various mRNA therapeutic properties of interest, such as **half life**, **ribosomal load**, and **translation rate** from **sequence only**.
3. We demonstrate the ability to conditionally guide **mRNAutilus** generation using property prediction regressors to **simultaneously design UTRs for and perform codon optimization on existing open reading frame templates**. Monte Carlo Tree Guidance shows consistent improvement of mRNA fitness over the generation time course, greatly outperforming existing wild-type mRNAs in *in-silico* evaluation.

Paper