mDD-0: mRNA Discrete Diffusion for Generation of Stable mRNA Sequences

Authors: Alyssa Morrow, Michal Jastrzebski, Jake Wintermute

Contributors: Siqi Zhao, Justin Gardin, Lood van Niekerk, Valentin Zulkower, Elise Flynn, Joshua Moller, Porfirio Quintero Cadena, Hao Shen, Dana Merrick, Ankit Gupta, Seth Ritter

INTRODUCTION

Recent advances in AI have enabled remarkable progress in the design and optimization of messenger RNA (mRNA) for RNA vaccines and therapeutics. While recent work on mRNA design has largely focused on designing individual components of the mRNA sequence [1,2,3], designing end-to-end mRNA sequences presents a unique challenge. Specifically, designing functional mRNA requires joint optimization of the coding sequence (CDS) for a protein of interest in addition to the 3' and 5' untranslated regions (UTRs).

Recently, diffusion models have emerged as a powerful generative framework for protein design [15,16] and sequence generation [17], for example, Ginkgo's recently released antibody discrete diffusion model. Building on these advances, we introduce mRNA discrete diffusion (mDD-0), a discrete diffusion model for the generation of mRNA sequences. The mDD-0 model is trained using genomic sequence data from hundreds of species as well as proprietary synthetic data. We show that mDD-0 can unconditionally generate mRNA sequences with similar sequence traits and predicted function to genomic sequences. We further demonstrate that, when paired with custom data generation, mDD-0 can optimize key functional features of an mRNA sequence, such as mRNA stability. The mDD-0 model outperforms conventional design strategies, such as genetic algorithms on a variety of metrics including predicted function, diverse candidate generation, and diversity from the training set.

A version of mDD-0 trained on genomic data only is accessible through Ginkgo's Model API and additional documentation is available. Ginkgo offers enhanced versions of this model and high-throughput data generation capabilities to fine-tune mDD-0 for your therapeutic payload of interest: **contact us today**.

OVERVIEW OF MRNA DISCRETE DIFFUSION (mDD-0)

mDD-0 was designed to learn mRNA sequence features across different species. The model implements a unique architecture to jointly learn from multimodal inputs: 3' UTRs, 5' UTRs, amino acid sequences and their corresponding coding sequences (CDS). Because we incorporate training data from hundreds of species when training our model, we condition each mRNA sequence on its species of origin, allowing for the generation of species-specific coding sequences during mRNA design.

Figure 1 demonstrates the model architecture for Ginkgo's mDD-0 model. The model takes as input a 3' and 5' UTR, an amino acid sequence and a species token. It outputs a 3' and 5' UTR and corresponding protein coding sequence.

During training, 3' and 5' UTR inputs are masked at various rates and mDD-0 learns to denoise each masked sequence. Amino acid sequence inputs are not masked, but instead are translated to a species-conditioned CDS in the model output. This allows the model to learn codon usage for each species.

After training, one can use mDD-0 to sample synthetic mRNA sequences by providing an amino acid sequence, species, and fully or partially masked 3' and 5' UTRs to design novel, synthetic mRNA sequences.

The architecture consists of different modules for each sequence input. The 3' and 5' UTRs are passed through BERT large language embedding modules that were first pre-trained on mammalian UTRs from 125 species. The Amino acid sequences use a pretrained ESM2-150M [4] as an embedding module, which is frozen during training. The species is encoded using a simple embedding layer. Embeddings for species, amino acid sequence, and both UTRs are then concatenated and passed through a lightweight transformer to produce a joint representation. In total, mDD-0 has 250M parameters.

DATASET PREPARATION

To collect mRNA sequences across multiple species, we curated gene annotation and genomic DNA for 324 vertebrate genomes from the Ensembl genome database [5]. Transcripts were filtered to include only protein-coding sequences. 5' and 3' UTRs were parsed from relative positions in gene annotations. Sequences with invalid start and stop codons, invalid coding sequence lengths, or missing UTRs were removed. After filtering, 7,168,632 mRNA sequences remained. Sequences were clustered using mmseqs2 [6] and distinct clusters were divided into training, validation, and test sets.

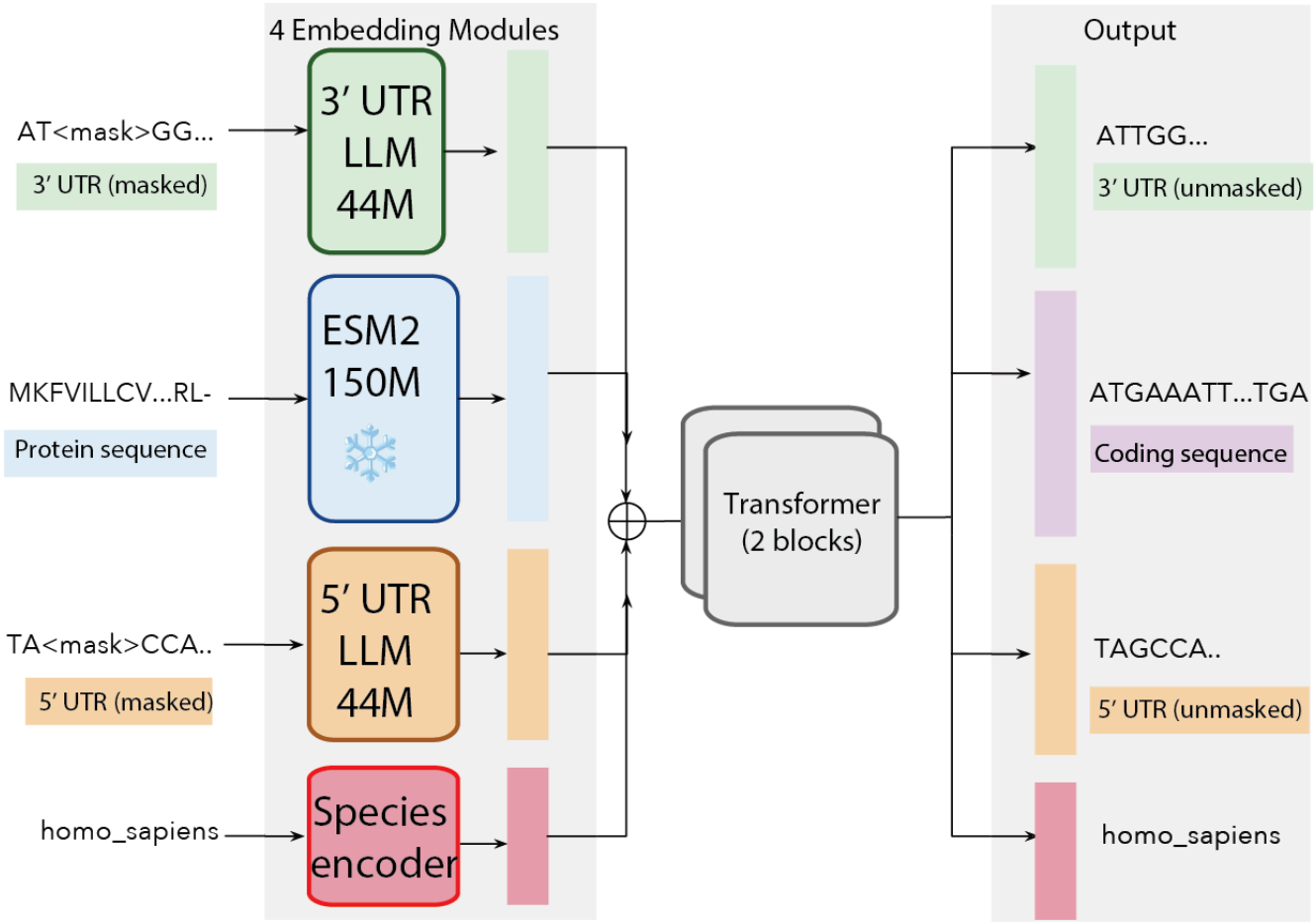

FIGURE 1. Architecture and training procedure from the mRNA discrete diffusion (mDD-0) model. mDD-0 contains four embedding modules. The 3' UTR, 5' UTR, amino acid sequence, and species for a given mRNA are passed through their respective modules to calculate embeddings. Embeddings are concatenated and passed through a lightweight transformer to learn the joint distribution across mRNA sequences and species. 3' and 5' UTRs are masked during training, and the model estimates unmasked nucleotides. Amino acid sequences are passed through ESM2-150M, whose weights are frozen, and are translated to its native coding sequence as model output.

MDD-0-GENERATED MRNA SEQUENCES RESEMBLE NATURALLY OCCURRING MRNA SEQUENCES

We first sought to understand how mDD-0 can be used to generate full mRNA sequences, and whether the coding sequences and UTRs generated using our model were similar to genomic sequences.

To evaluate mDD-0, we first sampled full mRNA sequences by masking whole 3' and 5' UTRs of genomic sequences, providing only the amino

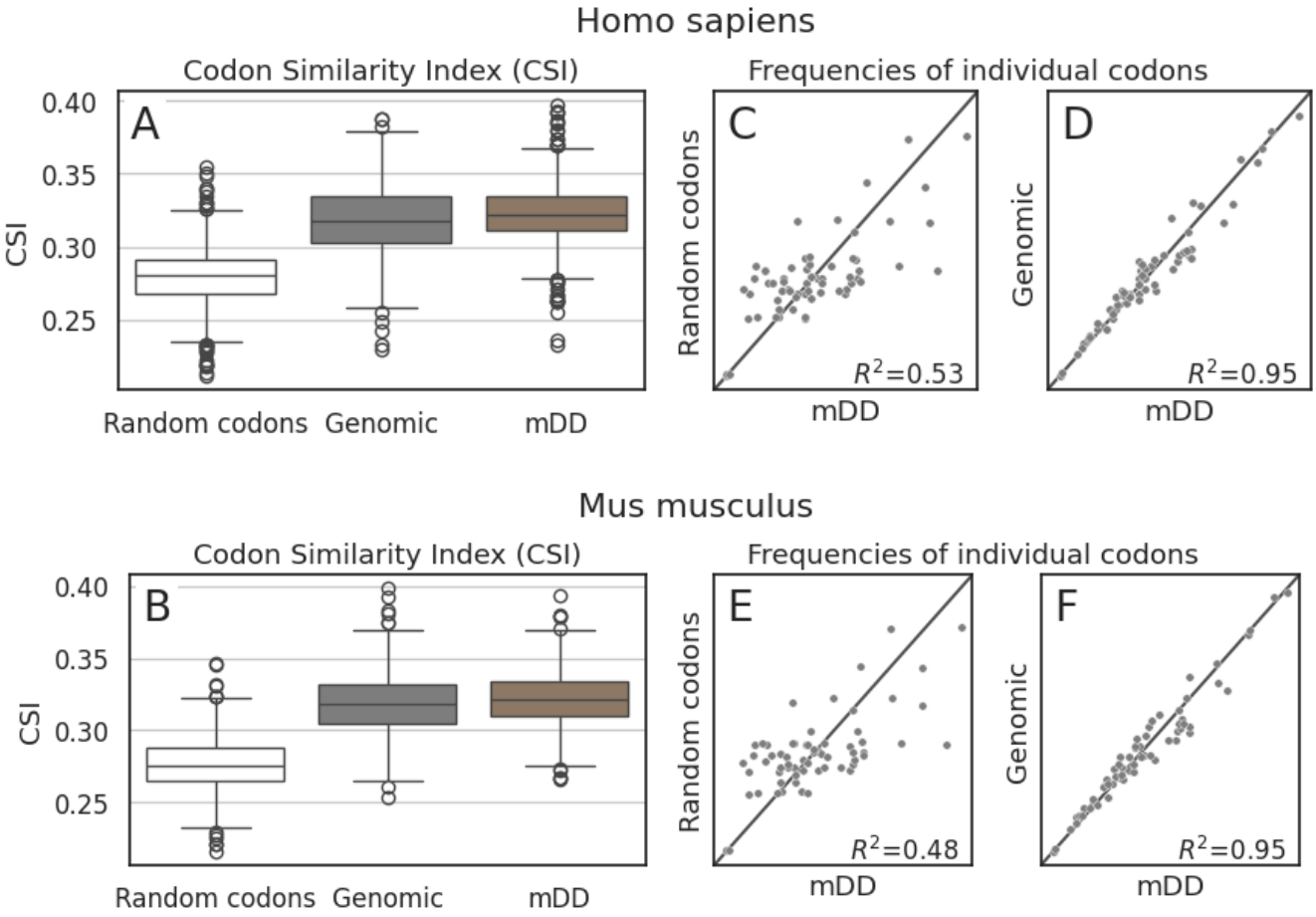

Figure 2. The mDD model predicts codon usage patterns in human and mouse transcripts. Codon Similarity Index (CSI) computed using (**A**) homo sapiens and (**B**) mus musculus genes for genomic CDSs and mDD-0-generated CDSs. As a control, we include the CSI of CDSs generated using random codon selection. For each species, CDSs were generated by conditioning on an amino acid sequence and species. CSI was calculated based on species-specific codon tables from [18]. (**C-F**) Comparison of codon frequencies between CDSs generated by mDD-0 or by random codon selection for (**C**) homo sapiens (R^2=0.53), and (**E**) mus musculus (R^2=0.48); and CDSs generated by mDD-0 or obtained from the genome for (**D**) homo sapiens: (R^2=0.95), and (**F**) mus musculus (R^2=0.95).

acid sequence and species as input to the model.

We then evaluated the quality of generated coding sequences (CDS) using 1212 and 818 amino acid sequences from *Homo sapiens* and *Mus musculus* (mouse) genomes, respectively, that were held out from training. We evaluate CDSs using a metric called the Codon Similarity Index (CSI) [7], which measures how similar codon usage is between a given gene and the rest of the genes for a target species.

Figure 2A,B demonstrates that the CSI of CDSs generated with mDD-0 is similar to the CSI of native, genomic CDSs. Additionally, Figures 2D,F demonstrate the codon frequencies of mDD-0 generated CDSs are similar to their native genomic sequences for *Homo sapiens* and *Mus musculus*. In contrast, Figures 2C,E demonstrate that codon frequencies of mDD-generated CDSs are less similar to CDS generated using random codon selection. These results indicate that a diffusion model can reliably sample CDSs similar to the codon usage of a specific species when conditioned on species and an amino acid sequence of interest.

We next evaluated 3' and 5' UTRs generated with mDD-0. UTRs were fully masked when input to the model, while the species and amino sequence were provided as a prompt to mDD-0.

We first observed that the GC content of the 3' and 5' UTRs generated using mDD-0 is similar to the GC distribution of genomic UTRs (Figure 3). In addition, the GC content of each UTR type exhibits a stronger overlap with its corresponding genomic GC content distribution. For instance, the generated 5' UTRs overlap more closely with genomic 5' UTRs than with genomic 3' UTRs (Figure 3B).

We additionally evaluated UTRs generated by the diffusion model to determine if critical

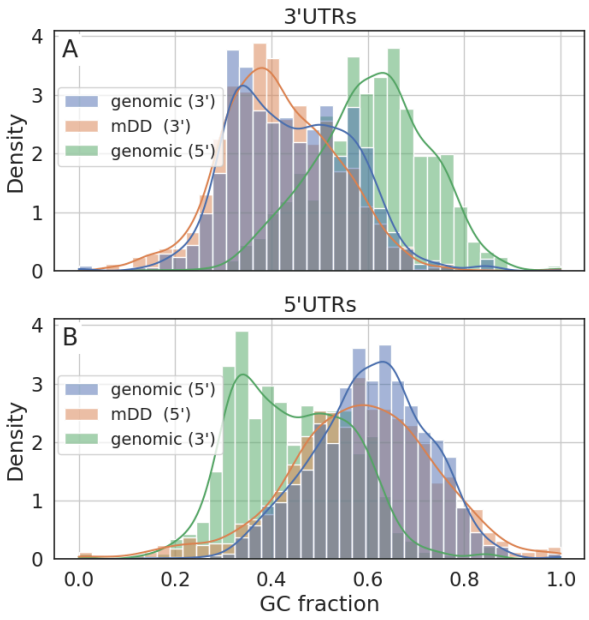

Figure 3. Distributions of GC content for 1,212 genomic human UTRs and UTRs sampled from mDD-0. (**A**) GC content for genomic and mDD-0 generated 3' UTRs. (**B**) GC content for genomic and mDD-0 generated 5' UTRs.

sequence features were retained in the UTRs. As an example, we found that the Kozak sequence, which signals the start of protein translation in eukaryotic mRNA [8], was placed in similar positions in generated 5' UTRs as in genomic sequences (Figure 4A). With the exception of four 5' UTRs generated with the diffusion model, Kozak occurrences in the diffusion-generated and genomic 5' UTRs had the same frequencies (Figure 4B). Additionally, we found that neither generated 5' UTRs nor genomic 5' UTRs contained upstream ORFs, as instances with upstream ORFs were filtered out from the training data.

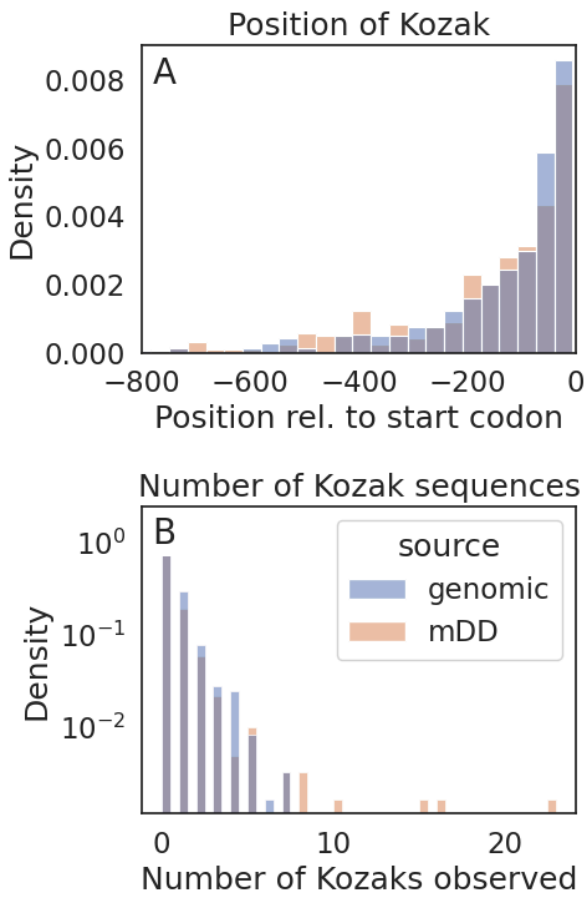

Figure 4. Distribution of Kozak sequence (GCCRCC) location and number of Kozak occurrences in the 5' UTR for 1212 genomic (human) and mDD-0 generated 5' UTRs. (**A**) The relative position of Kozak sequences found in genomic and generated 5' UTRs. Position is relative to the upstream start codon. (**B**) The number of Kozak sequences observed in each genomic and generated 5' UTR.

To assess generated 5' UTRs, we used a model trained on ribosomal load measurements from [3, 9] to predict the ribosomal load (RBL) of genomic and mDD-generated 5' UTRs. We first trained a 5' UTR RBL model on data from [3] and [9] and found our model to be highly performant on a held-out test set of 5' UTRs (Spearman = 0.87). Using this predictive model, we find that the predicted RBL of mDD-0 generated and genomic human 5' UTRs are similar (Figure 5A).

We similarly assessed 3' UTRs by comparing the predicted stability of genomic and generated 3' UTRs. To predict stability, measured as z-score normalized half-life of mRNA, we leveraged our predictive model described in [1], which predicts mRNA stability based on the 3' UTR sequence in the Hek293T cell line. We further refer to our 3' UTR stability model as Delphi. Similar to 5' UTRs, we found that the predicted stability of mDD-generated 3' UTRs were similar to that of genomic sequences (Figure 5B). Together, these results demonstrate that mDD-0 can generate CDSs and UTRs that are similar in sequence traits and predicted function to that of genomic sequences.

MRNA DISCRETE DIFFUSION CAN BE GUIDED TO OPTIMIZE FUNCTIONAL TRAITS OF THE MRNA SEQUENCE

While mDD-0 can generate CDSs and UTRs similar to genomic sequences, we ideally want to generate mRNA sequences with unique traits that improve existing mRNA sequences' stability, immunogenicity, and expression. Figure 5 demonstrates that even though UTRs

designed with mDD-0 have similar predicted stability and ribosomal load to genomic sequences, many UTRs have been observed in the training data that have higher ribosomal load and stability than human genomic sequences. We want to generate sequences that are not just similar to genomic sequences, but are also optimized for the traits we care about.

Additionally, one burning question we had is whether large generative models are worth the effort. We previously found that genetic algorithms, which iteratively mutate and recombine known stable 3' UTRs while using a predictive model for filtering mutants, designed 3' UTRs that translated experimentally *in vivo*, while maintaining higher diversity than alternative design strategies such as mutagenesis [1]. Therefore, we were curious how mRNA sequences generated using a generative model like mDD-0 compared to a simpler method, such as our experimentally validated genetic algorithm.

To evaluate whether mDD-0 can generate mRNA sequences that optimize specific traits, we utilized our supervised 3' UTR model that predicts mRNA stability (Delphi), [1], to guide mDD-0 to generate mRNAs with 3' UTRs that increase the stability of the mRNA construct. Specifically, we generate stable 3' UTRs using the same 5' UTR and CDS present in the mRNA construct used in training data for

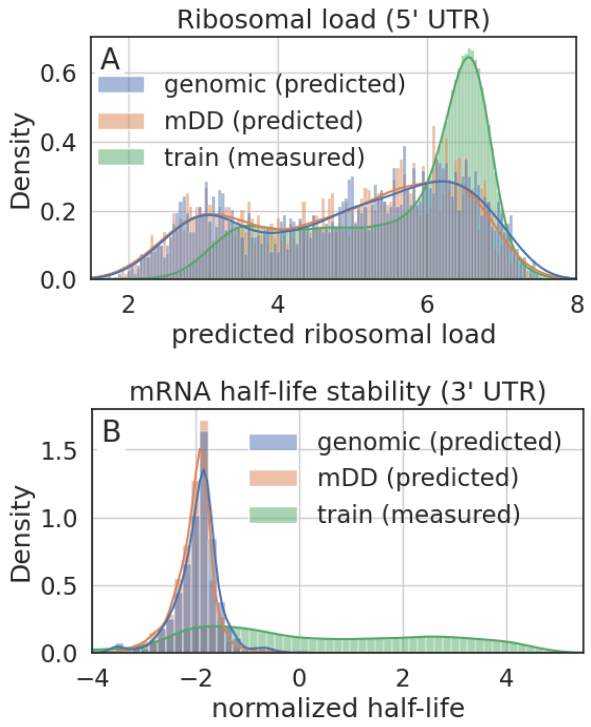

Figure 5. Predicted ribosomal load and mRNA stability (z-score normalized half-life) of mDD-generated UTRs are similar to genomic UTRs. (**A**) Predicted ribosomal load (RBL) of 1212 mDD-0 generated 5' UTRs and genomic 5' UTRs used as templates for generation (human), and training dataset used to train the ribosomal load model. (**B**) Predicted stability of 1212 genomic (human) and mDD-0 generated 3' UTRs. "train" represents all genomic and synthetic 3' UTRs that were included in the training data for Delphi. Delphi is a predictive model for 3' UTR stability introduced in [1].

Delphi to align mDD-0 generation with the data used to train the guiding predictive model.

We coupled mDD-0 with two algorithms for guidance: direct preference optimization (DPO) [10] and Soft Value-based Decoding in

Diffusion models (SVDD), [11]. Briefly, we use DPO to fine-tune mDD-0 by curating pairs of ranked 3' UTR sequences (ranked by stability predictions, or experimental measurements, when available). Pairs of ranked 3' UTRs are used to teach the model to preferentially generate 3' UTRs with higher mRNA stability, while down sampling 3' UTRs with lower stability. mDD-0 fine-tuned using DPO is then sampled from to generate stable 3' UTRs. SVDD, on the other hand, guides sampling by incorporating a predictive model that scores intermediate sequences during the generation process, steering the generation toward higher-quality outputs while maintaining diversity. We used each method to generate 20,000 3' UTRs.

MRNA DISCRETE DIFFUSION CAN BE GUIDED TO GENERATE STABLE MRNAS

Figure 6 demonstrates the predicted stability of 3' UTRs that were generated with mDD-0 without using a predictive model (mDD, no guidance) or that were conditionally generated using SVDD to guide generations of stable 3' UTRs. We found that DPO and SVDD generated sequences had similar predicted stability (data not shown). For this reason, we further investigate SVDD due to its increased computational efficiency compared to DPO.

We additionally include 3' UTRs that we generated using our *in vivo* generation pipeline for 3' UTRs that uses Delphi and a genetic algorithm [1]. Although we observed that SVDD and DPO generated 3' UTRs with significantly higher predicted stability than unconditionally generated 3' UTRs, genetic algorithms could more easily converge on 3' UTR designs with higher predicted stability (Figure 6).

WHEN GENOMIC DATA IS NOT ENOUGH: mDD-0 FINE-TUNED ON SYNTHETIC DATA SIGNIFICANTLY IMPROVES STABILITY OF GENERATED MRNA SEQUENCES

One limitation of mDD-0 is that it was initially only trained on genomic sequences, constraining the generation of mRNA sequences to the known evolutionary space. However, Delphi is trained on a combination of genomic and synthetic stable 3' UTRs and thus diverges from the known set of evolutionary defined UTRs.

To this end, we fine-tuned mDD-0 on synthetic 3' UTRs that were experimentally validated to have high stability. Figure 6 shows that 3'UTRs generated using fine-tuned mDD and SVDD (mDD fine-tuned + SVDD) have significantly higher predicted stability compared to UTRs generated using the genomic-data-only mDD-0 model (mDD+SVDD). Using mDD fine-tuned, the top 3' UTRs exceed the predictive stability of our experimentally validated genetic algorithm. Additionally, the predicted stability for these generated sequences is not

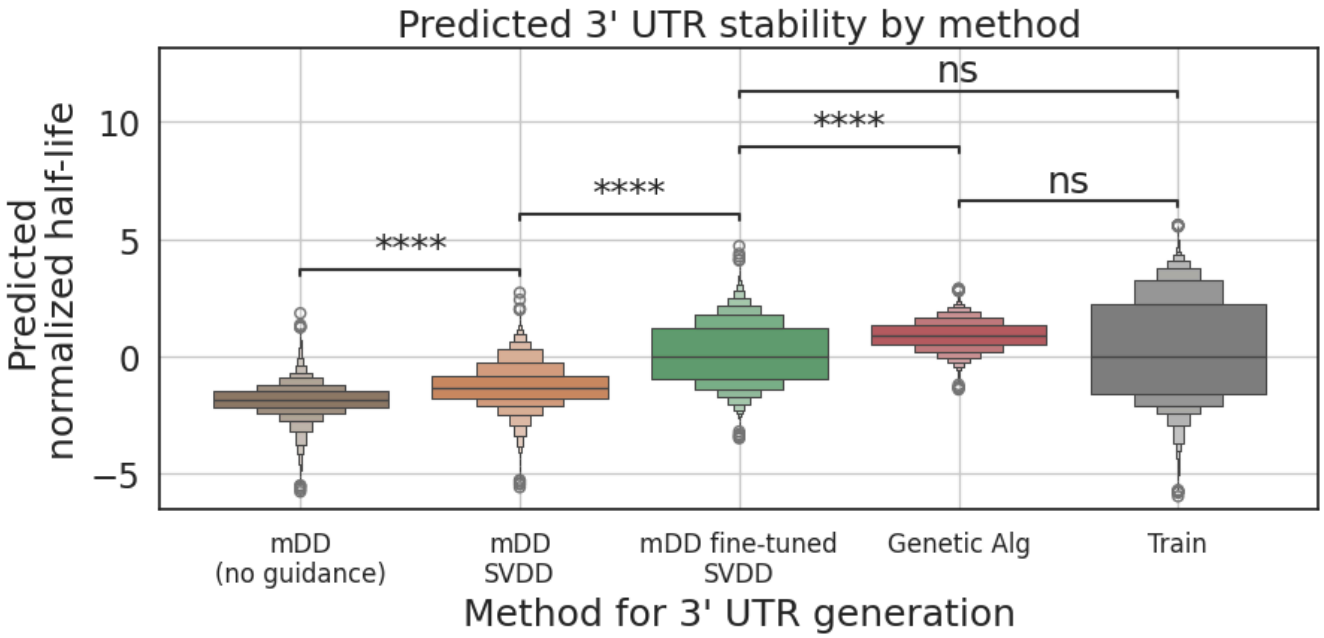

Figure 6. Predicted stability of generated 3' UTRs using four methods. mDD (no guidance) does not use Delphi to guide designs. mDD+SVDD, mDD fine-tuned + SVDD, and Genetic Alg (Genetic algorithm) methods use Delphi to guide designs of stable 3' UTR sequences. mDD only pre-trains on genomic mRNA sequences, while mDD fine-tuned was further trained on in vitro validated stable synthetic 3' UTRs. Train includes predictions for all training data used to train Delphi. Predicted stability was evaluated using an additional oracle trained with a different architecture than Delphi to avoid over-estimating the stability of UTRs that were hyperoptimized to Delphi's parameters. Significance was calculated using a one-sided Mann-Whitney test.

statistically different from the training data. These results suggest the importance of alignment between data used to train a generating distribution (mDD-0) and the predictive model used to guide the generating distribution.

mDD-0 GENERATED 3' UTRS MAINTAIN HIGH SEQUENCE DIVERSITY WHEN COMPARED TO A GENETIC ALGORITHM AND GENOMIC UTRS

Aside from predicted stability, we next sought to evaluate the diversity of generated 3' UTRs for all methods. Diversity of generated designs is particularly important for two reasons:

1. A set of diverse candidate sequences may have a higher chance of yielding a subset that translates from *in vitro* screens to animal models, as the underlying sequence features driving stability and expression may also be more diverse.

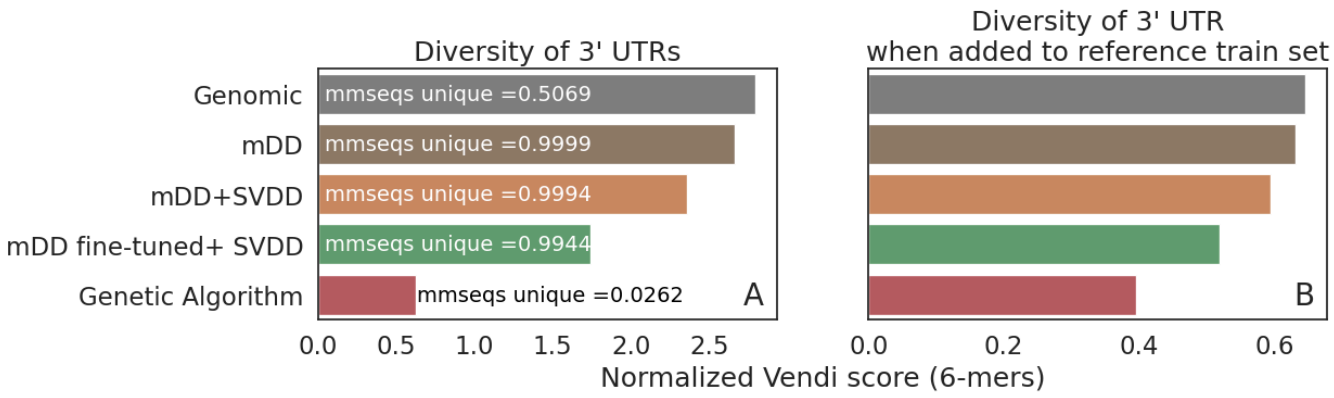

Figure 7. Vendi score, a metric of diversity, computed using 6-mer features from 3' UTRs. Lower scores indicate lower diversity. (**A**) Vendi score of 3' UTRs generated using four design methods. As a control, we include the diversity of genomic 3' UTRs from Delphi training data. Each bar is labeled with the number of clusters identified using mmseqs, normalized for the total number of sequences in each group (mmseqs unique). (**B**) Similar to (**A**), but combining each sequence set with a representative mmseqs subsampled set of the Delphi stability training data.

2. Experimental validation of diverse and novel designs will assist the next iteration of model training more than highly similar designs, which may result in the model converging too quickly to local optima during design, missing interesting areas of the sequence space that could be explored [12].

To evaluate (1), we analyzed the diversity of generated 3' UTRs, computed using the Vendi Score [13] on k-mers of length 6. We found that regardless of the algorithm chosen for design, 3' UTRs generated using mDD-0 had significantly higher diversity than sequences generated with a genetic algorithm (Figure 7A).

We additionally used mmseqs to compute the number of unique clusters in each group of generated 3' UTRs printed in (Figure 7A). We find that 3' UTRs generated with mDD-0 have the highest number of clusters, when normalized for the total number of sequences in each group. Generated sequences with a genetic algorithm consisted of only 525 unique clusters out of a total of 20,000 sequences (0.026 cluster fraction), with the biggest cluster containing 1313 sequences. In contrast, mDD-0 generated UTRs contained high clustering diversity (>0.99 mmseqs cluster fraction). These clustering metrics indicate superior diversity of 3' UTRs generated with mDD-0.

To assess diversity for (2), we computed the overall diversity of designed sequences when combined with the training data for Delphi. This metric gives us a relative idea of how adding generated 3' UTRs to our existing mRNA stability training data would increase or

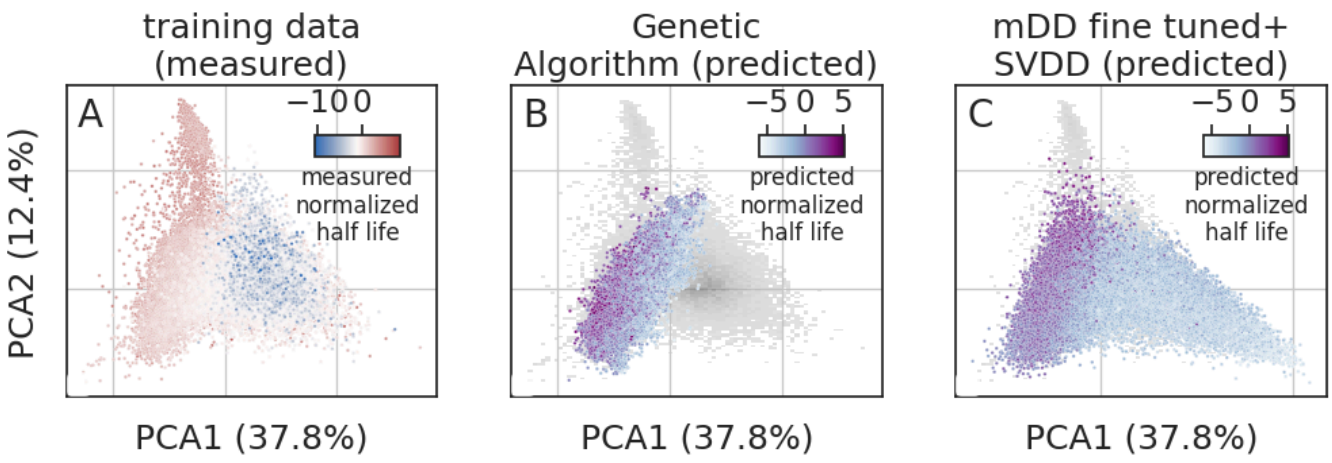

Figure 8. Principal component analysis (PCA) of 3' UTRs. PCA was run on model embeddings collected using Delphi. The first two components are visualized. (**A**) PCA of a representative subset of 3' UTRs used to train Delphi. UTRs are colored by measured z-score normalized half-lives. (**B**) PCA of UTRs generated with a genetic algorithm. UTRs are colored by predicted z-score normalized half-lives. Grey points represent training data. (**C**) PCA of UTRs generated with SVDD and mDD-0 fine-tuned on synthetic 3' UTRs. UTRs are colored by predicted z-score normalized half-lives.

decrease diversity for training our next iteration of models. As a control, we compute the diversity of adding distinct mmseqs clustered genomic sequences to the reference train set, which maintains the highest overall Vendi diversity (Figure 7B). We found that 3' UTRs designed with a genetic algorithm decreased the overall diversity more than 3' UTRs generated using mDD-0 (Figure 7B). We additionally observe lower diversity of UTRs generated with the fine-tuned mDD-0 model, suggesting convergence of generated sequences to a subset of the training data used to train Delphi (Green, Figure 7B).

Figure 8 visually demonstrates the sequence space searched by both the genetic algorithm and SVDD run with mDD-0 fine-tuned on synthetic 3' UTRs. We collected model embeddings from Delphi for a representative set of training data and for generated 3' UTRs. Figure 8 visually demonstrates the vast sequence space that SVDD searches when compared to a genetic algorithm and the original training data.

Together, these results suggest the superiority of sequences generated with mDD-0 in terms of (1) diverse candidate generation and (2) diversity from the training set, when compared to a genetic algorithm.

mDD-0 GENERATED 3' UTRS HAVE SIMILAR SEQUENCE FEATURES TO DELPHI TRAINING DATA

We next wanted to understand whether any of our methods for generating 3' UTRs had a tendency to generate functionally disrupted UTRs that may not transfer experimentally. To assess this *in silico*, we sought to measure how different generated UTRs were from the experimentally validated training data that was used to train Delphi. Similar to work from [14], we train XGBoost classifiers to empirically test whether generated UTRs can be discriminated from a sampled subset of sequences from the Delphi training set (the reference set). As a control, we include a non-overlapping sample of 3' UTRs from the train set that should be non-differentiable from the reference set.

Figure 9 demonstrates that 3' UTRs generated using mDD-0 fine-tuned on synthetic data are the hardest to differentiate from a representative set of the 3' UTRs used to train Delphi (Figure 9, auPRC = 0.73). As a control, we include a set of 3' UTRs from the training set that were not included in the representative set, which are expected to be the hardest to differentiate from the representative set (Figure 9, auPRC = 0.72). However, sequences designed with a genetic algorithm are the easiest to distinguish from the reference train set (auPRC = 0.93). While *in silico* metrics cannot replace experimental data, these preliminary metrics indicate the sequences designed using a genetic algorithm are further from the sequence distribution of measured stable 3' UTRs, suggesting potential pathological 3' UTRs.

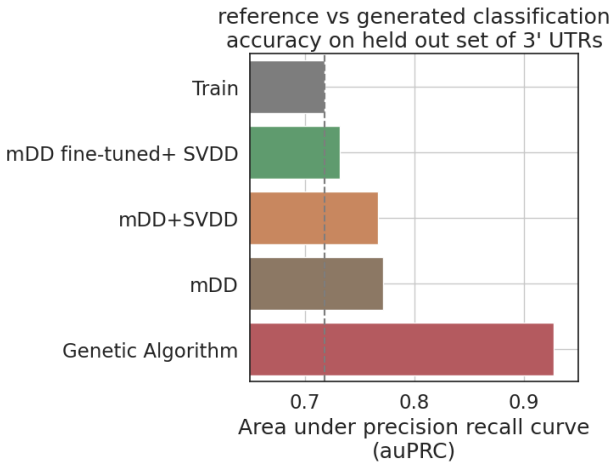

Figure 9. xGBoost classifiers were trained on 6-mers from generated 3' UTRs and reference 3' UTRs subsampled from Delphi training data. An xGBoost classifier was trained to classify sequences as "generated" or "from the reference set" for each set of generated sequences. A held-out set of 3' UTRs was used to evaluate the auPRC of the classifier's ability to differentiate between generated and reference 3' UTRs. Higher auPRC indicates larger representations of sequences features that differentiate generated and reference 3' UTRs, suggesting a deviation between sequence distributions.

CONCLUSION

Taken together, these results show that mDD-0 can capture not just the general sequence features of genomic mRNA sequences, but can also be guided to design synthetic mRNA sequences that achieve desired properties for mRNA therapeutics.

We have seen that mDD-0 fine-tuned on synthetic data is able to generate 3' UTRs with high predicted stability. When compared to a genetic algorithm, these sequences are (1) more diverse, (2) maintain higher diversity when added to a train set for future design iterations, and (3) are less differentiable from the training set.

While these results demonstrate optimized design of the 3' UTR using a predictive oracle trained on a specific payload, SVDD and mDD-0 can be easily extended to all components of the mRNA sequence. When coupled with Ginkgo data generation capabilities, we can construct predictive models of mRNA stability, translation rate, and protein expression that are specific to your payload of interest and use these models to guide the design of mRNA sequences with mDD-0.

Access to mDD-0 trained on genomic data is available through the Ginkgo Model API. Sign up for our API here. Read the documentation for using mDD-0 with the Ginko API here. To learn more about Ginkgo's newest mRNA diffusion model trained on both genomic and proprietary synthetic data, **contact us today**.

REFERENCES

1. Morrow et al. ML-driven design of 3' UTRs for mRNA stability. NeurIPS 2024 Workshop on AI for New Drug Modalities (2024).

2. Fallahpour et al. CodonTransformer: a multispecies codon optimizer using context-aware neural networks. bioRxiv 2024.09.13.612903 (2024).

3. Sample, P.J., Wang, B., Reid, D.W. et al. Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. Nat Biotechnol 37, 803–809 (2019).

4. Lin et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction bioRxiv 2022.07.20.500902; doi: https://doi.org/10.1101/2022.07.20.500902

5. Harrison et al. Ensembl 2024. Nucleic Acids Res. (2024), 52(D1):D891–D899.

6. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. 35, 1026–1028 (2017).

7. Sabath N, Wagner A, Karlin D. Evolution of viral proteins originated de novo by overprinting. Mol Biol Evol. 2012 Dec;29(12):3767-80. doi: 10.1093/molbev/mss179.

8. Kozak M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell. 1986 Jan 31;44(2):283-92. doi: 10.1016/0092-8674(86)90762-2. PMID: 3943125.

9. Karollus A, Avsec Ž, Gagneur J (2021) Predicting mean ribosome load for 5'UTR of any length using deep learning. PLoS Comput Biol 17(5): e1008982. https://doi.org/10.1371/journal.pcbi.1008982

10. Rafailov et al. Direct Preference Optimization: Your Language Model is Secretly a Reward Model (2024). https://arxiv.org/abs/2305.18290

11. Li et al. Derivative-Free Guidance in Continuous and Discrete Diffusion Models with Soft Value-Based Decoding (2025) .https://openreview.net/forum?id=2fgzf8u5fP

12. de Boer, C.G., Taipale, J. Hold out the genome: a roadmap to solving the cis-regulatory code. Nature 625, 41–50 (2024). https://doi.org/10.1038/s41586-023-06661-w

13. Dan Friedman and Adji Bousso Dieng. The Vendi Score: A Diversity Evaluation Metric for Machine Learning. 2023. URL: https://arxiv.org/abs/2210. 02410.

14. Avantika Lal, Laura Gunsalus, Anay Gupta, Tommaso Biancalani, Gokcen Eraslan. Polygraph: A Software Framework for the Systematic Assessment of Synthetic Regulatory DNA Elements. doi: https://doi.org/10.1101/2023.11.27.568764

15. Ingraham, John B., Max Baranov, Zak Costello, Karl W. Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier et al. Illuminating protein space with a programmable generative model. Nature 623, no. 7989 (2023): 1070-1078.

16. Wu, Kevin E., Kevin K. Yang, Rianne van den Berg, Sarah Alamdari, James Y. Zou, Alex X. Lu, and Ava P. Amini. Protein structure generation via folding diffusion. Nature communications 15, no. 1 (2024): 1059.

17. Zehui Li, Yuhao Ni, William A V Beardall, Guoxuan Xia, Akashaditya Das, Guy-Bart Stan, Yiren Zhao. DiscDiff: Latent Diffusion Model for DNA Sequence Generation. https://arxiv.org/abs/2402.06079 (2024).

18. Nakamura, Y., Gojobori, T. and Ikemura, T. Codon usage tabulated from the international DNA sequence databases:status for the year 2000. (2000) Nucleic Acids Res. 28, 292.