HELIX-mRNA: A HYBRID FOUNDATION MODEL FOR FULL SEQUENCE mRNA THERAPEUTICS

Matthew Wood
Helical
Luxembourg
matthew@helical-ai.com

Mathieu Klop
Helical
Luxembourg
mathieu@helical-ai.com

Maxime Allard [*]
Helical
Luxembourg
maxime@helical-ai.com

ABSTRACT

mRNA-based vaccines have become a major focus in the pharmaceutical industry. The coding sequence as well as the Untranslated Regions (UTRs) of an mRNA can strongly influence translation efficiency, stability, degradation, and other factors that collectively determine a vaccine's effectiveness. However, optimizing mRNA sequences for those properties remains a complex challenge. Existing deep learning models often focus solely on coding region optimization, overlooking the UTRs. We present Helix-mRNA, a structured state-space-based and attention hybrid model to address these challenges. In addition to a first pre-training, a second pre-training stage allows us to specialise the model with high-quality data. We employ single nucleotide tokenization of mRNA sequences with codon separation, ensuring prior biological and structural information from the original mRNA sequence is not lost. Our model, Helix-mRNA, outperforms existing methods in analysing both UTRs and coding region properties. It can process sequences 6x longer than current approaches while using only 10% of the parameters of existing foundation models. Its predictive capabilities extend to all mRNA regions. We open-source the model (https://github.com/helicalAI/helical) and model weights (https://huggingface.co/helical-ai/helix-mRNA).

1 INTRODUCTION

Engineered messenger Ribonucleic Acids (mRNAs) have emerged as a versatile platform for advanced therapeutics and vaccines, offering rapid design, scalable production, and flexible administration Pardi et al. (2018). The swift deployment of COVID-19 mRNA vaccines exemplified these advantages, reducing conventional vaccine development timelines from years to months Park et al. (2021). Beyond infectious disease, mRNA-based immunotherapies have shown durable anti-tumour efficacy in cancer Sahin et al. (2017). Further, optimizing mRNA for stability and translation efficiency enables high-yield protein expression in both microbial and mammalian hosts Eisenhut et al. (2020). These advances underscore the broad applicability of mRNA engineering for biomedical and industrial needs, ranging from gene therapies to large-scale protein manufacturing. Deep learning methods have shown promise in further accelerating mRNA therapeutic and vaccine design Castillo-Hair et al. (2024), spanning traditional architectures to foundation models, which are models trained on large datasets that serve as building blocks for different downstream tasks Yazdani-Jahromi et al. (2024).

Current approaches face three key limitations that particularly impact mRNA sequence analysis. First, the short context length used by current approaches is problematic because mRNA sequences vary dramatically in length, from only a few nucleotides to several thousand, making it difficult to capture long-range dependencies effectively. Second, simplified tokenization schemes risk losing important biological structures that exist in the original sequences Yazdani-Jahromi et al. (2024). Third, existing models like HELM Yazdani-Jahromi et al. (2024), CodonBERT Ren et al. (2024b), and Optimus 5-Prime Castillo-Hair et al. (2024) are specialised for specific mRNA regions, making them inflexible and inefficient to adapt to other mRNA regions without complete retraining.

---

[*]Corresponding Author

To address these limitations, we introduce Helix-mRNA, a hybrid (attention-based Vaswani et al. (2023) and state-space-based Gu et al. (2022); Gu & Dao (2024)) foundation model capable of capturing full-length mRNA sequences. We employ a two-stage pre-training method that addresses task-specific specialisation challenges. Our approach features single nucleotide tokenization with a structural codon representation. Helix-mRNA outperforms current models across various downstream benchmarks. These tasks include prediction of translation efficiency, stability and degradation from UTR and codon information.

2 METHODS

We break down the different components of Helix-mRNA and how each of them overcome current challenges in the mRNA modality. Helix-mRNA enables long sequence processing through a hybrid state-based and attention architecture, achieving single nucleotide resolution and still maintaining precise coding region representation. Further, we use a two stage pre-training approach through Warmup-Stable-Decay (WSD) scheduling Hu et al. (2024), allowing for a balance between generalisation and specialisation.
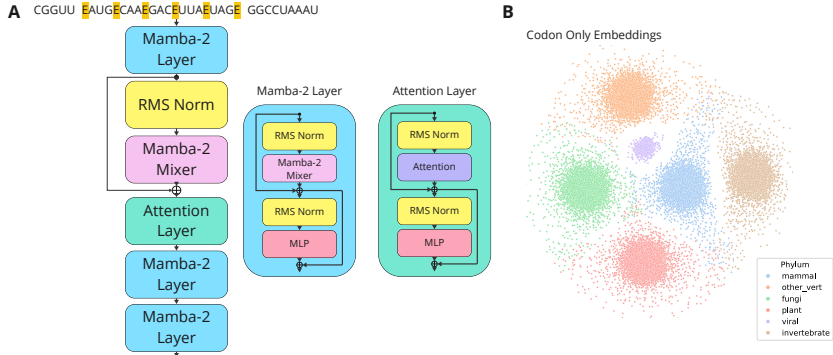
2.1 HYBRID ARCHITECTURE

Figure 1: **A)** Helix-mRNA hybrid architecture incorporating state-based and attention-based approaches. We show how we retain coding region structure with an additional token highlighted in the figure. **B)** Embeddings from initial pre-training, generated using only coding regions from mRNA sequences not seen during training.

We train a 5.19 million parameter model, 10% of the parameters of Transformer HELM Yazdani-Jahromi et al. (2024), with a hybrid SSM and attention-based architecture shown in Figure 1 taking inspiration from Nemotron Parmar et al. (2024), Jamba Lieber et al. (2024), Zamba Glorioso et al. (2024) and Samba Ren et al. (2024a). The ratio of Mamba-2 layers, MLP layers, and attention layers was selected from the different ablation studies performed by Waleffe et al. (2024) with 44.4% Mamba-2 layers, 44.4% MLP layers and 11.1% attention layers. By preceding attention layers with Mamba-2 layers , the recurrent nature of the Mamba-2 layer encodes positional information needed by the attention layer Vaswani et al. (2023); Dao & Gu (2024). This architectural approach enables Helix-mRNA to read long mRNA sequences with high granularity, combining state-based long sequence handling with the in-context retention capabilities of attention-based methods.

2.2 SINGLE NUCLEOTIDE AND CODON ENCODING

Long mRNA sequences are traditionally computationally challenging, addressed by analysing subregions of the sequence Yazdani-Jahromi et al. (2024) and coarser tokenization methods that group nucleotides together Ren et al. (2024b). With our hybrid architecture enabling long sequence lengths, we tokenize full-length sequences at single nucleotide resolution. We map each base (*A, C, U,* and *G*) to a unique integer, ensuring the preservation of complete biological information without any data loss. A key focus of our approach is the mRNA coding region, which plays a critical role in

translation efficiency, protein expression, and mRNA stability Boël et al. (2016), Hanson & Coller (2018). To emphasize the importance of codon structure and enhance pattern extraction, we introduce a special character E to clearly denote codon separation. We pre-train on a maximum input length of 12288 tokens, 6x the context length of Transformer HELM Yazdani-Jahromi et al. (2024).

2.3 DIVERSE mRNA SEQUENCES ACROSS MULTIPLE PHYLA

We use a taxonomically diverse pre-training dataset which was specifically curated to capture both deep evolutionary conservation patterns in eukaryotic sequences and the distinct nucleotide compositions and structural characteristics of viral genomes. The broad phylogenetic coverage enables the model to learn robust representations of sequence features across different evolutionary distances and genomic architectures. We gathered all RefSeq mRNA sequences[1] from a diverse set of eukaryotes, including vertebrates (mammals and non-mammals), plants, invertebrates, fungi, and 238 clinically relevant viruses O'Leary et al. (2016) to pre-train our Helix-mRNA model (See Appendix A.2).

2.4 TWO-STAGE SPECIALISED PRE-TRAINING

We utilise a two stage autoregressive unsupervised pre-training strategy, guided by the WSD learning rate scheduler Hu et al. (2024). In the first stage, the WSD scheduler manages a warm-up phase followed by a stable learning rate, allowing the model to process data of varying quality, creating a general and robust base model. In the second stage, the WSD scheduler transitions into its decay phase, where training focuses exclusively on high-quality data. During the second phase, human only mRNA sequences are used to refine the model for human specific tasks. This two stage process enables efficient learning from mixed-quality data in the initial stage and achieving specialised distillation through the targeted second stage.

3 RESULTS

Table 1: Benchmark comparisons showing Helix-mRNA, CodonBERT, Transformer HELM and Transformer XE. Spearman rank correlations are shown for all tasks.

Task	CodonBERT	HELM	XE	Helix-mRNA
MLOS Flu Vaccines	0.54	0.70	0.65	**0.79**±0.121
mRFP Expression	0.77	0.85	0.82	**0.86**±0.008
mRNA Stability/iCodon	0.35	**0.53**	0.50	0.52±0.004
Tc-Riboswitch	0.50	0.63	0.57	**0.64**±0.033
Vaccine Degradation	0.78	0.83	0.80	**0.84**±0.032

After the first stage of pre-training, we use Uniform Manifold Approximation (UMAP) to visualise the embeddings of unseen coding sequences from each of the different phyla used during pre-training. Figure 1 shows that Helix-mRNA clearly clusters coding sequences from different phyla through unsupervised pre-training alone. We demonstrate that despite pre-training on full sequences, Helix-mRNA can extract relevant information from different subregions without the need for further training. Viruses cluster very clearly and separately from all other eukaryotic phyla in the embedding space. This distinct separation implies that the model has learned biologically meaningful features, such as differences in codon usage and sequence structure, enabling it to discriminate viral from non-viral sequences.

3.1 BIOLOGICAL DOWNSTREAM TASKS

We first evaluate Helix-mRNA on coding region related property prediction tasks including stability, degradation and translation efficiency. Our results demonstrate that Helix-mRNA outperforms existing approaches like CodonBert, Transformer HELM, and Transformer XE across multiple bench-

---

[1]https://ftp.ncbi.nih.gov/refseq/release/

marks, including an E. coli related tasks even though Helix-mRNA was not pre-trained on any prokaryotic data, further highlighting generalisability to various tasks (See Appendix A.4. We assessed performance using the following benchmarks where we use a 70% training split, a 15% validation split and a 15% test split:

- mRNA Stability/iCodon : Stability profiles from vertebrates Diez et al. (2022).
- mRFP Expression : Protein production levels in E. coli variants Nieuwkoop et al. (2023).
- MLOS dataset: Haemagglutinin antigen encodings for flu vaccines Ren et al. (2024b).
- Tc-Riboswitches : Tetracycline riboswitch dimers with switching factor measurements Groher et al. (2018).
- SARS-CoV-2 Vaccine Degradation : mRNA sequences optimized for structural features Leppek et al. (2022).

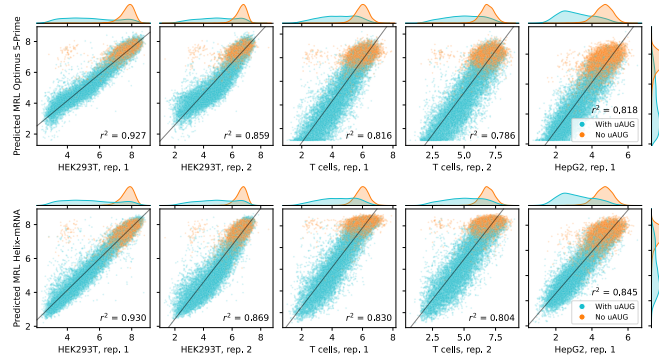## 3.2 UTR5 MEAN RIBOSOME LOAD PREDICTION

Figure 2: Benchmark comparison between Helix-mRNA and Optimus 5-Prime on task specific fine-tuning to predict Mean Ribosome Load (MRL) across 3 cell lines (HEK293T, T cells, and HepG2) using two replicates, reproduced from the Optimus 5-Prime codebase released with the paper Castillo-Hair et al. (2024). Results show the r^2 correlation values between the predicted MRL and the true MRL.

The benchmarks presented in Table 1 focus solely on the coding region. However, since Helix-mRNA is capable of utilising both UTR regions and coding regions, we evaluate the model on tasks specific to the 5' UTR. We evaluated our model's predictive capabilities by replicating the tasks from Castillo-Hair et al. (2024), predicting Mean Ribosome Load (MRL) across three cell lines (HEK293T, T cells, and HepG2) using two replicates, shown in Figure 3.2. Fine-tuning for 5 epochs enabled us to outperform Optimus 5-Prime on MRL prediction tasks, with the greatest improvements seen in T cells and HepG2 cell lines, where Optimus 5-Prime struggles the most. By applying our model to their established tasks for cell line specific MRL prediction, we demonstrate our approach's performance and generalisability to different downstream tasks, further highlighting our model's versatility.

4 CONCLUSION

In this work we presented Helix-mRNA, a hybrid architecture combining attention mechanisms and SSMs, taking inspiration from recent advances in long-context natural language models Waleffe et al. (2024), allowing for longer context lengths and better in-context retention capabilities. The extra context length combined with single-nucleotide and codon structure tokenization allows the model to capture additional features useful in mRNA sequence design, namely codons, UTR regions, and secondary structures. We show that our tokenization methods doesn't impact performance and allows us to use the model in coding region only tasks, outperforming Transformer HELM Yazdani-Jahromi et al. (2024), Transformer XE and CodonBERT Ren et al. (2024b) or in UTR 5' region

specific tasks by outperforming Optimus 5-Prime Castillo-Hair et al. (2024). By effectively merging state-of-the-art sequence modelling techniques with two-stage pre-training, Helix-mRNA paves the way for more robust, efficient, and broadly applicable mRNA therapeutics in both clinical and industrial domains.

REFERENCES

Grégory Boël, Reka Letso, Helen Neely, W Nicholson Price, Kam-Ho Wong, Min Su, Jon D Luff, Mayank Valecha, John K Everett, Thomas B Acton, et al. Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature*, 529(7586):358–363, 2016.

Sebastian Castillo-Hair, Stephen Fedak, Ban Wang, Johannes Linder, Kyle Havens, Michael Certo, and Georg Seelig. Optimizing 5'UTRs for mRNA-delivered gene editing using deep learning. *Nature Communications*, 15(1):5284, 2024.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality, 2024. URL https://arxiv.org/abs/2405.21060.

Michay Diez, Santiago Gerardo Medina-Muñoz, Luciana Andrea Castellano, Gabriel da Silva Pescador, Qiushuang Wu, and Ariel Alejandro Bazzini. iCodon customizes gene expression based on the codon composition. *Scientific Reports*, 12(1):12126, 2022.

Peter Eisenhut, Aman Mebrahtu, Mona Moradi Barzadd, Niklas Thalén, Gerald Klanert, Marcus Weinguny, Anna Sandegren, Chao Su, Diane Hatton, Nicole Borth, et al. Systematic use of synthetic 5'-utr rna structures to tune protein translation improves yield and quality of complex proteins in mammalian cell factories. *Nucleic Acids Research*, 48(20):e119–e119, 2020.

Paolo Glorioso, Quentin Anthony, Yury Tokpanov, James Whittington, Jonathan Pilault, Adam Ibrahim, and Beren Millidge. Zamba: A Compact 7B SSM Hybrid Model, 2024. URL https://arxiv.org/abs/2405.16712.

Ann-Christin Groher, Sven Jager, Christopher Schneider, Florian Groher, Kay Hamacher, and Beatrix Suess. Tuning the performance of synthetic riboswitches using machine learning. *ACS synthetic biology*, 8(1):34–44, 2018.

Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces, 2024. URL https://arxiv.org/abs/2312.00752.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces, 2022. URL https://arxiv.org/abs/2111.00396.

Gavin Hanson and Jeff Coller. Codon optimality, bias and usage in translation and mRNA decay. *Nature reviews Molecular cell biology*, 19(1):20–30, 2018.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies, 2024. URL https://arxiv.org/abs/2404.06395.

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

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avashalom Manevich, Nir Ratner, Noam Rozen, Erez Shwartz, Mor Zusman, and Yoav Shoham. Jamba: A Hybrid Transformer-Mamba Language Model, 2024. URL https://arxiv.org/abs/2403.19887.

Thijs Nieuwkoop, Barbara R Terlouw, Katherine G Stevens, Richard A Scheltema, Dick De Ridder, John Van der Oost, and Nico J Claassens. Revealing determinants of translation efficiency via whole-gene codon randomization and machine learning. *Nucleic acids research*, 51(5):2363–2376, 2023.

Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2016.

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

Jung Woo Park, Philip NP Lagniton, Yu Liu, and Ren-He Xu. mRNA vaccines for COVID-19: what, why and how. *International journal of biological sciences*, 17(6):1446, 2021.

Jupinder Parmar, Shrimai Prabhumoye, Joseph Jennings, Mostofa Patwary, Sandeep Subramanian, Dan Su, Chen Zhu, Deepak Narayanan, Aastha Jhunjhunwala, Ayush Dattagupta, Vibhu Jawa, Jiwei Liu, Ameya Mahabaleshwarkar, Osvald Nitski, Annika Brundyn, James Maki, Miguel Martinez, Jiaxuan You, John Kamalu, Patrick LeGresley, Denys Fridman, Jared Casper, Ashwath Aithal, Oleksii Kuchaiev, Mohammad Shoeybi, Jonathan Cohen, and Bryan Catanzaro. Nemotron-4 15b technical report, 2024. URL https://arxiv.org/abs/2402.16819.

Liliang Ren, Yang Liu, Yadong Lu, Yelong Shen, Chen Liang, and Weizhu Chen. Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling. *arXiv preprint arXiv:2406.07522*, 2024a.

Zilin Ren, Lili Jiang, Yaxin Di, Dufei Zhang, Jianli Gong, Jianting Gong, Qiwei Jiang, Zhiguo Fu, Pingping Sun, Bo Zhou, and Ming Ni. CodonBERT: a BERT-based architecture tailored for codon optimization using the cross-attention mechanism. *Bioinformatics*, 40(7):btae330, 05 2024b. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae330. URL https://doi.org/10.1093/bioinformatics/btae330.

Ugur Sahin, Evelyna Derhovanessian, Matthias Miller, Björn-Philipp Kloke, Petra Simon, Martin Löwer, Valesca Bukur, Arbel D Tadmor, Ulrich Luxemburger, Barbara Schrörs, et al. Personalized rna mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*, 547 (7662):222–226, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, 2023. URL https://arxiv.org/abs/1706.03762.

Roger Waleffe, Wonmin Byeon, Duncan Riach, Brandon Norick, Vijay Korthikanti, Tri Dao, Albert Gu, Ali Hatamizadeh, Sudhakar Singh, Deepak Narayanan, Garvit Kulshreshtha, Vartika Singh, Jared Casper, Jan Kautz, Mohammad Shoeybi, and Bryan Catanzaro. An Empirical Study of Mamba-based Language Models, 2024. URL https://arxiv.org/abs/2406.07887.

Mehdi Yazdani-Jahromi, Mangal Prakash, Tommaso Mansi, Artem Moskalev, and Rui Liao. HELM: Hierarchical Encoding for mRNA Language Modeling, 2024. URL https://arxiv.org/abs/2410.12459.

6

A APPENDIX

A.1 SPECIFIC ARCHITECTURE IMPLEMENTATION DETAILS

We use a hidden dimension of 256 throughout the model with SiLU activation functions. For the Mamba-2 layers, we employ a head dimension of 32, a single group, and a kernel size of 4 for convolution, with an expansion factor of 2. The attention implementation uses Flash Attention 2 with 32 attention heads and 8 key-value heads. For MLP layers, we maintain an intermediate size of 512, representing an expansion ratio of 2. No biases are used in the attention and MLP layers. The model consists of 9 total layers, with a layer configuration of 4 Mamba-2 layers, 4 MLP layers and a single attention layer in the configuration $M+M*+M+M+$, where M represents a Mamba-2 layer, + represents an MLP layer and * represents attention layers.

Table 2: Model parameters used for Helix-mRNA.

Params (M)	No. Layers	Model Dim	Attn. Heads	State Dim.	No. Groups	Pos. Emb.	Seq. Len
5.19	8	256	16	128	1	None	12288

Table 3: Pre-training Details.

No. Tokens (B)	Pre-Training Time (Days)	No. Devices	Device Type
332	2.3	32	NVIDIA A100-SXM4-40GB

A.2 PRE-TRAINING DATASETS

All of the datasets can be found under the NCBI ftp server: `https://ftp.ncbi.nih.gov/refseq/release/`. We downloaded all the files ending by '.rna.gbff.gz' and extracted the utr and coding sequences via our pre-processing pipeline. Out of the 56 Million sequences we subsampled 27 million sequences due to budgetary constrains to train our model.

The viral component encompassed the main human pathogens, including respiratory viruses (SARS-CoV-2, influenza A/B/C, RSV), retroviruses (HIV-1/2, HTLV-1/2), hepatotropic viruses (HBV, HCV), herpesviruses (HSV-1/2, EBV, VZV), and emerging arboviruses (Zika, Dengue 1-4). The dataset consists of 27 million sequences, with class contributions being, 37.6% other vertebrates (`https://ftp.ncbi.nih.gov/refseq/release/vertebrate_other/`), 24.4% mammals (`https://ftp.ncbi.nih.gov/refseq/release/vertebrate_mammalian/`), 22.8% invertebrates (`https://ftp.ncbi.nih.gov/refseq/release/invertebrate/`), 13.7% fungi (`https://ftp.ncbi.nih.gov/refseq/release/fungi/`) and 1.4% viruses (`https://ftp.ncbi.nih.gov/refseq/release/viral/`).

A.3 FINE-TUNING

We unfreeze the 2 last layers when fine-tuning on downstream tasks, while freezing the rest of the parameters.

A.4 EVALUATION DETAILS

Due to the HELM code not yet being publicly available, we referenced their published research findings directly instead of independently reproducing their experimental methodology in Table 1 Yazdani-Jahromi et al. (2024). For our analysis, we selected the highest performance values from either the Masked Language Modelling or Causal Language Modelling variants of the Transformer XE and Transformer HELM models. The authors of HELM report the best results achieved for the

tasks, with the exception of the MLOS Flu Vaccines task, where the average performance across three random splits is reported due to the absence of predefined splits. To align with their methodology, we adopt a similar approach but also include an error margin to highlight the variability of results across five runs.

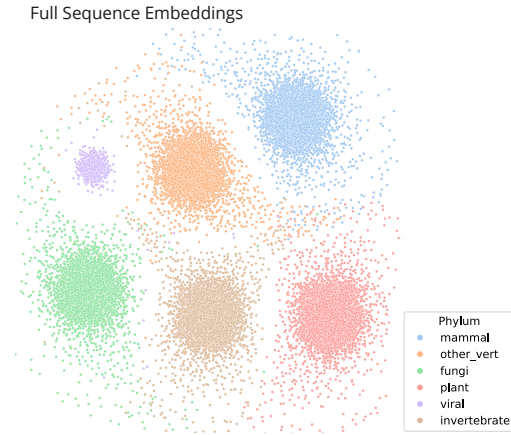## A.5 FULL SEQUENCE EMBEDDINGS AFTER PRE-TRAINING

Figure 3: Helix-mRNA embeddings from initial pre-training, generated using full mRNA sequences including both the coding and untranslated regions.

Embeddings presented in Figure A.5 show Helix-mRNA's ability to process full length sequences not limited to just the coding region.