

A 5' UTR language model for decoding untranslated regions of mRNA and function predictions

In the format provided by the
authors and unedited

Table of Contents

Table of Contents	1
A. Data Description	2
A.1. Pre-processing the data.....	2
A.2. Data of five species from the Ensembl database.....	2
A.3. Randomized 50-nucleotide oligomer data with mean ribosome load.....	2
A.4. Human 5' UTR data for translation efficiency.....	3
A.5. Dataset for the Internal Ribosome Entrance Site.....	4
A.6. In-house UTR design and experimental validation.....	4
A.6.1. 5' UTR sequence design by genetic algorithm.....	4
A.6.2. Plasmid synthesis and in vitro transcription.....	5
A.6.3. Cell culture.....	5
A.6.4. mRNA transfection and luciferase Assay.....	5
B. Details of Training	6
B.1. Hyper-parameters of pre-training process.....	6
B.2. Three types of representations generated by pre-training model.....	6
B.3. Splitting strategies of training and test data for MRL tasks.....	7
B.4. Experimental settings of Independent test.....	7
B.5. Ablation study on the pretraining and finetuning strategy.....	8
B.6. Model size and runtime comparison.....	9
C. Related Work	10
D. Results	12
D.1. Performance metrics of UTR-LM across different tasks.....	12
D.2. Performance metrics of UTR-LM compared with benchmarks.....	13
D.3. Repeating the UTR-LM MRL with different random seeds for stability test..	14
E. Model Interpretation	16
E.1. UTR-LM embeddings for differentiating between species.....	16
E.2. Correlation between MFE and sequence length.....	16
E.3. Attention Score Analysis and Biological Findings.....	17
E.4. Exhaustive scan of 3-12 mer fragments.....	19
References	21

A. Data Description

A.1. Pre-processing the data

Initially, we removed any coding sequence (CDS) or non-5' UTR fragments from the raw sequences. Next, we identified and removed duplicate sequences. Then, we truncated the sequences to fit within a range of 30 to 1022 bp, following the guidelines from previous work established by Cao et al.⁷ and ESM³⁸. Lastly, we filtered out incorrect or low-quality sequences following the work established by Sample et al.⁶ and Cao et al.⁷

A.2. Data of five species from the Ensembl database

We collected 214,349 unlabeled 5' UTR sequences from the Ensembl database³³, spanning five species: human, rat, mouse, chicken, and zebrafish. These species have high-quality and manual gene annotations. The potential benefits of incorporating 5' UTRs from an additional 309 species, which have comparatively lower-quality gene annotations, remain an avenue for future exploration. Detailed statistics for the data from these five species can be found in Supplementary Table 1.

After data cleaning, the counts of 5' UTRs from the Ensembl database³³ are 77,835, 27,740, 48,378, 31,577, and 28,819, respectively. We chose these species as they represent a diverse spectrum of vertebrates with varying evolutionary connections. The details can be found in Supplementary Table 1.

Supplementary Table 1. Endogenous data collected from the Ensembl database. The min and max of sequence lengths are 30 and 1022, respectively.

Species	Counts (#)	# With ATG	# Without ATG	Mean of Length	Median of Length
Human	77,835	48,128	29,707	249	190
Rat	27,740	17,639	10,101	309	201
Mouse	48,378	29,624	18,754	234	178
Chicken	31,577	20,999	10,578	233	193
Zebrafish	28,819	21,816	7,003	240	172
All	214,349	138,206	76,143	250	186

A.3. Randomized 50-nucleotide oligomer data with mean ribosome load

Sample et al.⁶ proposed eight distinct 5' UTR libraries, each containing random 50 nucleotide sequences, to evaluate translation rules using mean ribosome loading (MRL) measurements. These libraries are called as U_1 , U_2 , Ψ_1 , Ψ_2 , $m_1\Psi_1$, $m_1\Psi_2$,

mC-U₁, and mC-U₂. In this context, the U, Ψ , m₁ Ψ , mC symbolize unmodified uridine, pseudouridine, and 1-methyl pseudouridine, and the coding sequence (CDS) of the mCherry, respectively. Specifically, the mC-U₁ and mC-U₂ libraries, containing about 200,000 5' UTRs each, are based on the mCherry CDS. The remaining libraries utilize the CDS of the enhanced green fluorescent protein (eGFP) and each comprises roughly 280,000 5' UTRs. Further details are provided in Supplementary Table 2.

Sample et al.⁶ generated 83,919 random 5' UTR sequences and 15,555 human 5' UTRs, covering a length range from 25 to 100 nucleotides. The number of total reads per 5' UTR in both libraries is greater than 10.

Supplementary Table 2. Eight randomized 50-nucleotide oligomer data with mean ribosome load.

Library	CDS	Uridine	Counts (#)	# With ATG	# Without ATG
U ₁	eGFP	Unmodified	280,000	151,059	128,941
U ₂	eGFP	Unmodified	280,000	150,114	129,886
mC-U ₁	mCherry	Unmodified	280,000	104,643	95,357
mC-U ₂	mCherry	Unmodified	280,000	103,975	96,025
Ψ ₁	eGFP	Pseudouridine	279,968	149,370	130,598
Ψ ₂	eGFP	Pseudouridine	280,000	149,648	130,352
m1 Ψ ₁	eGFP	1-methylpseudouridine	279,968	149,246	130,722
m1 Ψ ₂	eGFP	1-methylpseudouridine	280,000	147,908	132,092

A.4. Human 5' UTR data for translation efficiency

We adopted the endogenous human 5' UTR data analyzed by Cao et al.⁷, which involves matching public RNA-Seq and Ribo-Seq datasets. Data from three distinct cell lines/tissues, specifically human embryonic kidney 293T (HEK), human prostate cancer cell (PC3), and human muscle tissue (Muscle), were assembled into three separate endogenous datasets. The mRNA expression level (EL) was determined using RNA-seq RPKM (Reads Per Kilobase of transcript per Million mapped reads), and the translation efficiency (TE) for each transcript was calculated by dividing the Ribo-seq RPKM by the RNA-seq RPKM⁷. In alignment with the limitation of commercially available single-stranded DNA template biosynthesis⁷, a fixed 5' UTR length of 100-bp was chosen for algorithm training. Transcripts that fail to meet the minimum coverage criteria are excluded, specifically those with RNA-Seq RPKM < 5 and Ribo-Seq RPKM < 0.1 for the gene. The HEK, PC3, and Muscle datasets

comprised 14,410, 12,579, and 1,257 sequences respectively. More detailed statistics can be found in Supplementary Table 3.

Cao et al.⁷ collected three datasets of human endogenous 5' UTRs, measuring both mRNA translation efficiency (TE) and expression level (EL), sourced from publicly available matched RNA-Seq and Ribo-Seq datasets. These datasets consist of three human cell lines/tissues: (i) Human Embryonic Kidney 293T (HEK), (ii) Human Prostate Cancer Cell (PC3), and (iii) Human Muscle Tissue (Muscle), with sample sizes of 14,410, 12,579, and 1,257, respectively. For each 5' UTR, we adopted the 100-bp sequence upstream of the CDS, covering 29% of human 5' UTRs. Statistics are shown in Supplementary Table 3.

Supplementary Table 3. Three datasets of human endogenous 5' UTRs with translation efficiency and mRNA expression level.

Source	Number (#)	# With ATG	# Without ATG	Mean of Length	Median of Length
Muscle	1257	613	644	183	126
PC3	12,579	6,075	6,504	192	132
HEK	14,410	7,018	7,392	185	127

A.5. Dataset for the Internal Ribosome Entrance Site

We assembled a dataset of sequences, each labeled binary based on their function as the internal ribosome entry site (IRES). This dataset was sourced from several studies, including those by Weingarten-Gabbay et al.²⁶, IRESbase²⁷, IRESite²⁸, Rfam²⁹, and the training dataset from IRESpred¹⁴. To maintain data uniqueness, We removed entries that were redundant across the different sources. The finalized dataset contained 46,774 entries. It is important to highlight that this dataset encompasses sequences with diverse lengths, from 6 to 1,730 nucleotides. Additionally, the label distribution is extremely unbalanced, with 37,602 sequences labeled as non-IRESs and 9,172 identified as IRESs.

A.6. In-house UTR design and experimental validation

A.6.1. 5' UTR sequence design by genetic algorithm

All sequence evolutions started with randomized sequences. Over the course of 100 iterations, either a single base or two bases (with a 50% probability) were randomly chosen in each iteration and mutated. The fitness of the sequence was then evaluated. If the new sequence had a performance that was higher or closer to the target, it was retained; otherwise, the original sequence was kept. The fitness was assessed by Optimus re-trained on MRL and TE tasks. From multiple rounds of

sequence evolution, 211 sequences with high predicted MRL and TE values were selected.

A.6.2. Plasmid synthesis and in vitro transcription

The designed 5' UTRs were synthesized and inserted into a custom plasmid backbone, positioned before the codon-optimized coding sequence of the luciferase gene and after the T7 promoter. A DNA template for in vitro transcription (IVT) was produced via plasmid linearization using BspQI restriction enzyme. For the IVT reaction, we utilized the HiScribe T7 high-yield RNA synthesis kit (NEB). As a cap structure analog, we employed the 3'-O-Me-m⁷ G(5') ppp(5')AmG RNA cap (TriLink Biotech). And the DNA template was digested with DNase I (NEB), and the IVT mRNA was purified using the Monarch® RNA Cleanup Kit (NEB).

A.6.3. Cell culture

C2C12 mouse myoblasts (ATCC, CRL-1772) were cultivated in DMEM with 10% FCS and 1 mM sodium pyruvate. HepG2 human hepatocellular carcinoma (ATCC, HB-8065) were cultivated in DMEM with 10% FBS and 1% Pen-Strep. All cells were cultivated at 37°C and a CO₂ content of 5%.

A.6.4. mRNA transfection and luciferase Assay

Prior to day of transfection, C2C12 and HepG2, 20000 cells/well were cultured with 100 µl of complete DMEM growth medium seed in 96-well tissue culture plate. In the following day, cells with density about 50~80% confluent were transfected with 25, 12.5, 6.25, 3.125ng of mRNA per well using Lipofectamine™ MessengerMAX™ Transfection Reagent (Invitrogen) according to manufacturer's protocol. Then, cells were incubated at 37°C in a CO₂ incubator. In vitro luciferase expression was measured at 16h and 48h post transfection using the One-Glo Luciferase Assay System (Promega). The efficacy of translation of different mRNA constructs was calculated as fold change relative to the benchmark mRNA, NCA-7d-5'UTR¹⁷, which comes from a pair of optimized 5' and 3' UTRs. To simplify, we only used the results from the 6.25ng dose at 16h in the C2C12 cell line to evaluate the machine learning methods.

B. Details of Training

B.1. Hyper-parameters of pre-training process

Supplementary Table 4. Data for pre-training process. The pre-training tasks labeled as MN, SS, and MFE correspond to the randomly 15% masked token, secondary structure, and minimum free energy tasks, respectively. Meanwhile, the downstream tasks identified as MRL, TE, and EL refer to mean ribosome loading, translation efficiency, and mRNA expression level tasks, respectively.

Data Source	MN task	SS task	MFE task		
			MRL task	TE task	EL task
Ensembl Database	√	√			
Randomized 50-nucleotides library	√	√	√		
Human datasets from tissue/cell lines	√	√		√	√

B.2. Three types of representations generated by pre-training model

The per-token representation consists of a list of vectors, with each vector corresponding to a single token in the input sequence. This representation captures details about the token, such as its part-of-speech, syntactic dependencies, and other relevant features. Because these vectors are derived from applying an attention mechanism to the hidden states, they highlight the importance of each token for the final prediction.

The mean representation is a single vector derived by averaging all the per-token representations within a sentence, serving as a summary representation of the input sequence.

A distinct token, termed the [CLS]-token, is positioned at the beginning of the sequence. Its representation serves as a learned vector that encapsulates the entire input sequence and is suitable for sequence classification or regression. The proposed pre-training model integrates supervised information as auxiliary labels, aiming to derive a representation tailored to specific tasks. This representation not only captures the global information of the input sequence, but also extracts the hidden information of specific downstream tasks. Employing this representation proves beneficial in generating accurate predictions for downstream tasks.

We have opted for the [CLS]-token representation when assessing downstream tasks.

B.3. Splitting strategies of training and test data for MRL tasks

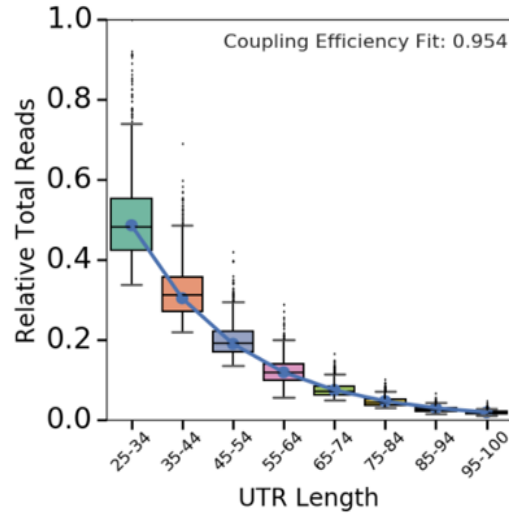
Sample et al.⁶ employed a read count-based sorting strategy, which we have termed Rank Split, to select 5' UTRs for training and testing. For the eGFP and mCherry libraries, the top 280,000 and 200,000 sequences were selected, respectively. From each library, the 20,000 UTRs with the highest read counts were allocated for the test set, with the rest reserved for training.

We introduced a method called Random Split to better capture data variability and reflect real-world situations. This method involves randomly selecting 20,000 sequences from the top 280,000 and 200,000 sequences in the eGFP and mCherry libraries, respectively, to form the test set, while the remaining sequences are used for training.

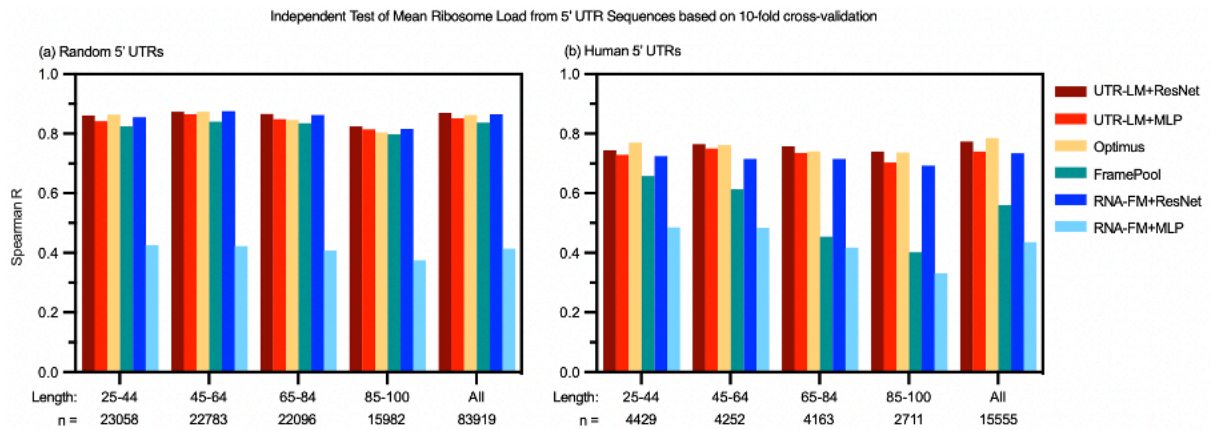
B.4. Experimental settings of Independent test

In the Optimus method⁶, a test set is formed from 100 5' UTRs of each length that have the deepest read coverage, ensuring approximately 10% of the library. This results in two test sets, each containing 7,600 5' UTRs—one composed of random sequences and the other of human sequences. The remaining 76,319 random 5' UTRs are used for model training. Both FramePool and RNA-FM use the same training set as Optimus, but their testing approaches differ. FramePool did not create a validation set and carried out model selection and evaluation using the 7,600 random 5' UTRs. On the other hand, RNA-FM employs the 7,600 random 5' UTRs as a validation set to determine optimal parameters and uses the 7,600 real human 5' UTRs as the test set. For fair comparison, we split the original training set ($n = 83,919$) into a training set and a validation set at a ratio of 9:1. These sets are used for model training and selection, respectively, with performance then evaluated on the two test sets, each containing 7,600 5' UTRs.

In the aforementioned experiments, 5' UTRs with the highest total read counts were used for the test set. Those 5' UTRs with more reads demonstrated higher resolution than those with fewer reads, and thus more accurately reflected their MRL (see Supplementary Fig. 1⁶). However, prediction methods should also perform well on test sets that are randomly split, as it can be challenging to guarantee high-quality samples in real-world situations. Thus, we conducted 10-fold cross-validation experiments for 83,919 random 5'UTR libraries and 15,555 human 5' UTR libraries, respectively. The results are displayed in Supplementary Fig. 2.



Supplementary Fig. 1. Fit the coupling efficiency on total reads in terms of 5' UTR length. This figure is proposed by Sample et al.⁶ The minima and maxima represent the lowest and highest data points excluding outliers, the centre refers to the median, the bounds of the box depict the first (25th percentile) and third quartiles (75th percentile), and whiskers extend to the most extreme data points not considered outliers. Percentiles correspond to the distribution points of the dataset. In this bar chart, each of the first seven bars represents 1000 samples for various UTR length intervals, while the last bar, corresponding to UTR lengths of 95-100bp, includes only 600 samples.



Supplementary Fig. 2. Performance for independent test sets across methods under 10-fold cross-validation for (a) 83,919 random 5'UTR libraries and (b) 15,555 human 5' UTR libraries.

B.5. Ablation study on the pretraining and finetuning strategy

We conducted additional experiments with an MN transformer (i.e., pre-trained language model) coupled with an MLP downstream predictor, using four distinct

approaches: (1) a MN transformer randomly initialized and weights freezed, with only downstream predictor fine-tuned, (2) a MN transformer with random initialization, where all weights were adjustable during finetuning, (3) a pre-trained MN transformer with freezed parameters, with only downstream predictor fine-tuned, and (4) a pre-trained MN transformer, where all parameters were tunable during fine-tuning (this is how we obtained the UTR-LM). The results are given in Fig. 2d and Fig. 3d.

In the MRL task, our UTR-LM model, fully fine-tuned based on the pretrained model, showed the highest performance (Spearman R: 0.962), significantly outperforming other methods (Fig. 2d). This demonstrates the effectiveness of both pretraining and comprehensive fine-tuning. The approach using a pretrained model with only the downstream predictor fine-tuned also performed well (0.6), indicating the value of a solid pretrained base. Models with random initialization, whether fully fine-tuned (0.4) or with only the downstream predictor fine-tuned (0.17), notably lagged behind, demonstrating the importance of pretraining. In the TE task across Muscle, PC3, and HEK datasets, a similar pattern emerges (Fig. 3d). These results highlight the contribution of both the fine-tuning and pre-training phase in our UTR-LM model. These new analyses provide a clearer understanding of the respective contributions of the pre-training and fine-tuning phases in our model.

B.6. Model size and runtime comparison

We calculated the model size of UTR-LM and recorded the runtime of fine-tuning the model on an P100 GPU for the MRL task on U_1 library (i.e., 280,000 sequences) and the TE task on the HEK dataset (i.e., 14,410 sequences). See Supplementary Table 5 for the results and comparison to other models. We notice that UTR-LM has <1.5m parameters, and it is only slightly more complex than FramePool and Optimus. Further, UTR-LM has only a fraction of model size compared to RNA-FM, making it a rather lightweight foundation model with superior performances.

Supplementary Table 5. Model size and runtime of fine-tuning UTR-LM, compared to other baselines.

Methods	Total number of parameters	Time needed for MRL task with U_1 library	Time for TE task with HEK dataset
UTR-LM-MLP	1,456,920	319s	28s
RNA-FM-MLP	99,547,227	418s	54s
RNA-FM-ResNet	99,709,995	528s	60s
RNABERT-MLP	514,777	427s	29s
RNABERT-ResNet	681,705	621s	40s
MTtrans	615,825	497s	29s
FramePool	282,625	12s	<1s
Optimus	474,681	17s	<1s

C. Related Work

Recent years have witnessed increasing interest in using language models (LM) to decode biological sequences³⁹. While there are plenty of protein LMs such as ESM-1b⁴⁰, TAPE⁴¹, ESMFold⁴², and ProGen⁴³, there are fewer LMs for learning semantic representations of DNA and RNA sequences. Benegas et al.⁴⁴ introduced the Genome pre-trained Network, an unsupervised model learning variant effects, gene structures, and DNA motifs in non-coding DNA. Dalla-Torre et al.⁴⁵ proposed the Nucleotide Transformer, a DNA sequence pre-training method that outperforms specialized methods in multiple prediction tasks using diverse species genomes. Ji et al.⁴⁶ developed DNABERT, pre-trained on the human genome, which effectively predicts promoter regions, splice sites, and transcription factor binding sites after fine-tuning. Yamada et al.³⁹ further demonstrated DNABERT's potential in RNA-related tasks, predicting RNA-binding protein interactions. Akiyama et al.¹² proposed RNABERT, which effectively utilizes the pre-trained BERT algorithm for non-coding RNA tasks and does not give results for any 5' UTR application. Chen et al.¹⁰ introduced RNA-FM, a model pre-trained on 27 million unannotated RNA sequences through self-supervision. Both RNABERT and RNA-FM have demonstrated utilities in structure prediction and function-related tasks. Models like SpliceBERT⁴⁷ and miProBERT⁴⁸ are pre-trained on specific RNA types, such as pre-mRNA and microRNA promoters. To the authors' best knowledge, no one has attempted to develop a language model specifically for modeling UTR sequences and their functions.

Several studies leverage massively parallel translation assays and specialized machine learning models, like ribosome profiling, to study 5' UTR variant effects on mRNA translation. Cao et al.⁷ developed random forests for predicting mRNA expression level (EL) and translation efficiency (TE) based on handcrafted features like *k*-mer frequencies, RNA folding energy, 5' UTR length, and the number of open reading frames. Sample et al.⁶ proposed Optimus, convolutional neural networks (CNN) for predicting the mean ribosome loading (MRL) of 5' UTRs, extendable to longer human 5' UTRs. Karollus et al.⁵ modified Optimus using global pooling, known as the FramePool method, to remove length restrictions. Chen et al.¹⁰ introduced the RNA-FM embedding, which was used in conjunction with a 32-layer residual network (ResNet) to predict MRL. Zheng et al.^{7,9} developed a multi-task CNN model for MRL and TE prediction using multiple data sources.

Several prior works studied machine learning models for identifying unannotated IRESes. IRESfinder¹⁵ is a logit model with framed *k*-mer features for cellular IRES prediction but may not be effective for viral IRES. DeepCIP¹⁶ is a deep learning model designed specifically for circRNA IRES prediction (a subset of all IRES), using both sequence and structure information. IRESPred¹⁴ is a support vector machine

model for predicting both viral and cellular IRES based on 35 handcrafted features including sequence properties, structural properties, and interaction probabilities between 5' UTR and small subunit ribosomal proteins¹³. IRESpy is an XGBoost model that is trained on 340 global *k*-mer features, minimum fold energy, and 32 features including sequence properties and structural properties.

D. Results

D.1. Performance metrics of UTR-LM across different tasks

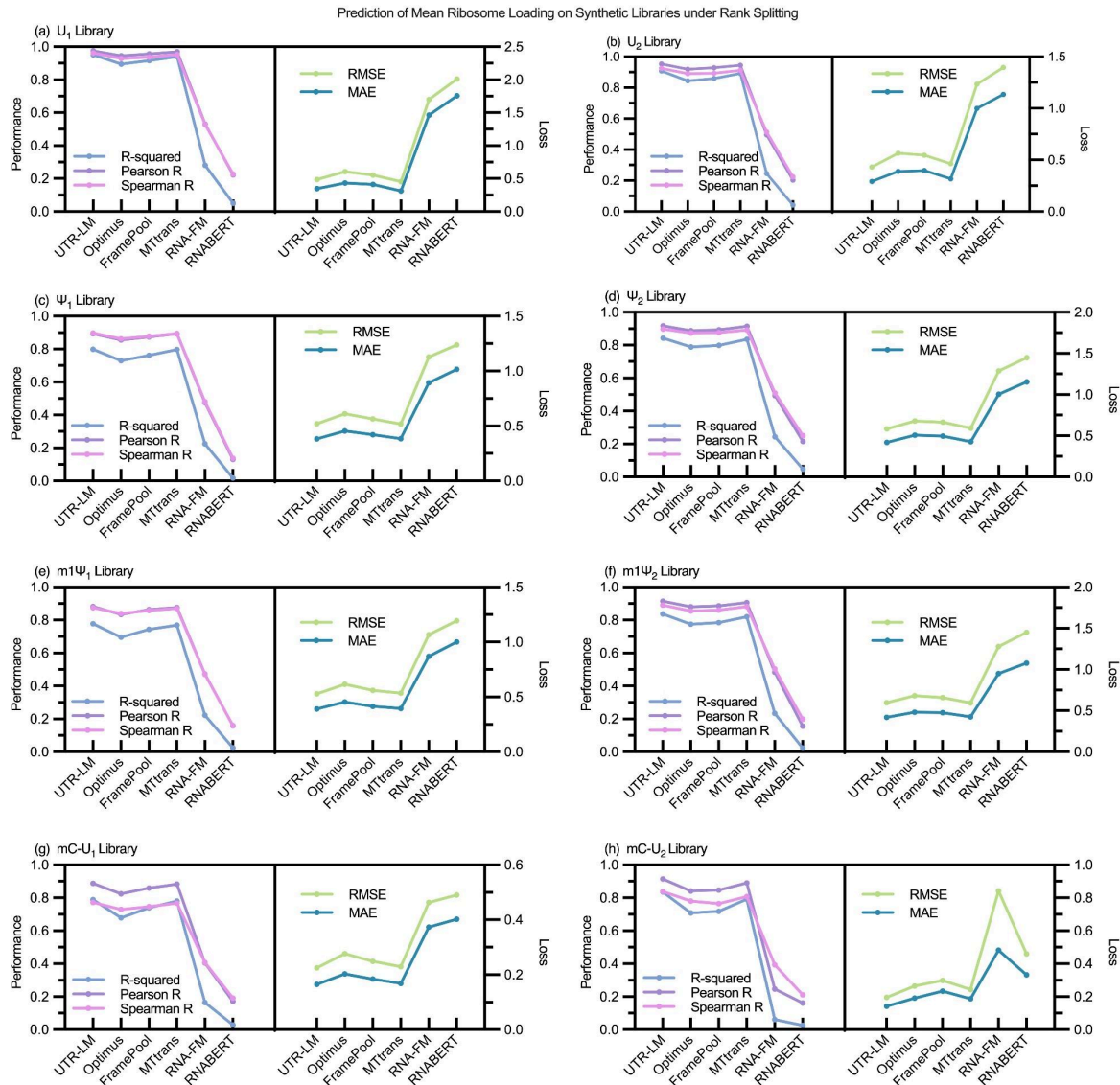
We report five performance metrics of the UTR-LM model across function prediction tasks in Supplementary Table 6.

Supplementary Table 6. All computed performance metrics of UTR-LM across different tasks.

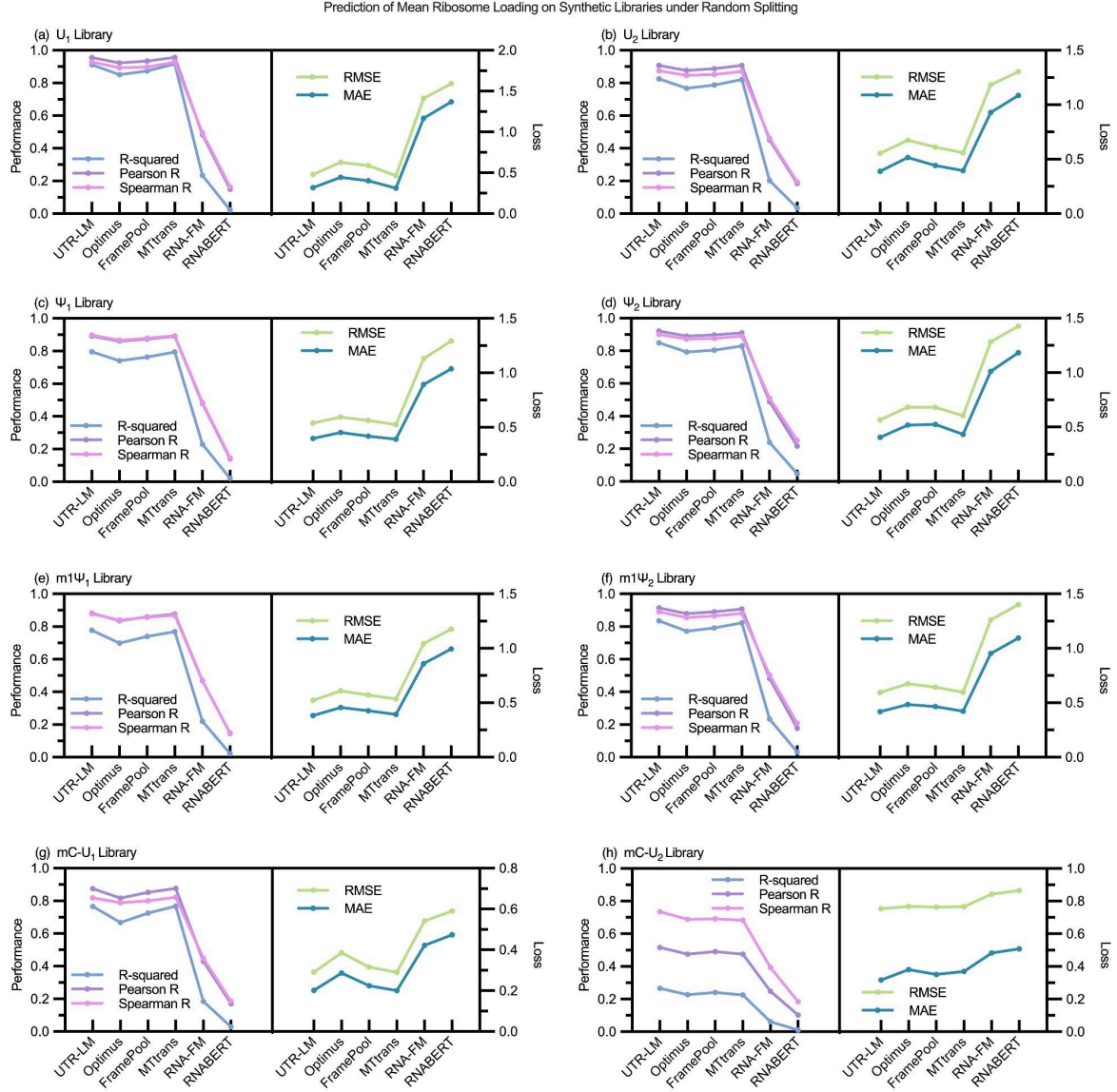
Task	Evaluation Strategy	Library/ Dataset	R-squared	Pearson R	Spearman R	RMSE	MAE
MRL	Held-out with Rank Splitting	U_1	0.951	0.975	0.962	0.487	0.348
		U_2	0.907	0.953	0.924	0.431	0.292
		Ψ_1	0.799	0.894	0.898	0.52	0.383
		Ψ_2	0.842	0.918	0.898	0.582	0.419
		$m1\Psi_1$	0.777	0.882	0.875	0.529	0.390
		$m1\Psi_2$	0.837	0.915	0.891	0.598	0.419
		$mC-U_1$	0.789	0.888	0.772	0.225	0.165
		$mC-U_2$	0.835	0.914	0.838	0.196	0.142
	Held-out with Random Splitting	U_1	0.912	0.955	0.928	0.48	0.319
		U_2	0.825	0.908	0.875	0.555	0.389
		Ψ_1	0.796	0.892	0.897	0.54	0.398
		Ψ_2	0.849	0.921	0.901	0.567	0.405
		$m1\Psi_1$	0.778	0.882	0.878	0.522	0.382
		$m1\Psi_2$	0.835	0.914	0.891	0.595	0.418
		$mC-U_1$	0.766	0.875	0.819	0.291	0.202
		$mC-U_2$	0.266	0.516	0.735	0.754	0.317
	Independent test	Random	0.790	0.889	0.870	0.755	0.548
		Human	0.734	0.856	0.801	0.702	0.517
TE	10-fold CV	Muscle	0.460	0.677	0.667	1.087	0.760
		PC3	0.469	0.685	0.646	0.748	0.509
		HEK	0.409	0.639	0.603	0.855	0.572
EL	10-fold CV	Muscle	0.504	0.707	0.661	0.919	0.601
		PC3	0.431	0.656	0.630	0.834	0.546
		HEK	0.469	0.685	0.651	0.744	0.486

D.2. Performance metrics of UTR-LM compared with benchmarks

We have expanded our evaluation to include additional metrics including Pearson R, R-squared, RMSE, and MAE. We computed all these performance metrics for UTR-LM as well as other baseline algorithms across subtasks; see Supplementary Fig. 3,4. We observe that UTR-LM is a consistent top performer across tasks regardless of the choice of metric.



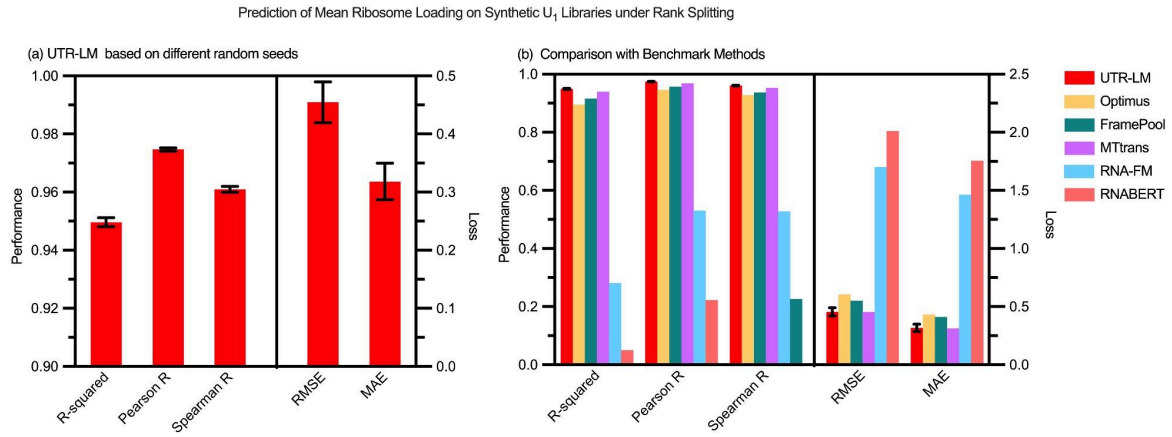
Supplementary Fig. 3. Model performances measured using 5 different metrics (R-squared, Pearson R, Spearman R, RMSE, MAE) across eight libraries of nucleotide 5' UTRs with Rank Splitting. UTR-LM exhibits consistently superior performances compared to other baselines, regardless of metric choice.



Supplementary Fig. 4. Model performances measured using 5 different metrics (R-squared, Pearson R, Spearman R, RMSE, MAE) across eight libraries of nucleotide 5' UTRs with Random Splitting. UTR-LM exhibits consistently superior performances compared to other baselines, regardless of metric choice.

D.3. Repeating the UTR-LM MRL with different random seeds for stability test

We experimented with varying random seeds to affirm the consistency and robustness of our UTR-LM model, focusing on the MRL prediction task. We repeated this experiment with ≥ 3 random seeds and reported performance metrics such as R-squared, Pearson R, RMSE, and MAE. See Supplementary Fig. 5 for these results. The figure showed that the UTR-LM consistently outperforms other baselines regardless of random seed and choice of metric. In addition, ANOVA analysis indicated that different random seeds do not lead to significant difference.



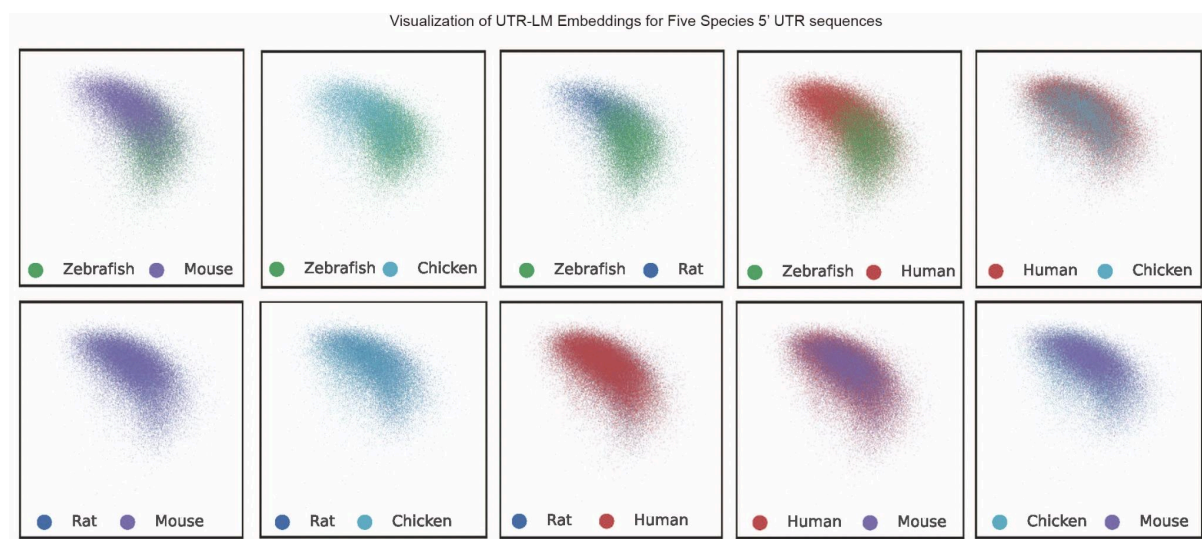
Supplementary Fig. 5. Stability analysis of UTR-LM with different random seeds. We report results obtained by repeating the mean ribosome loading prediction with random seeds/initialization and different metrics (using the U₁ library and rank splitting). Left figure shows performances of UTR-LM measured in different metrics with confidence intervals; Right figure shows the comparison against other models. Overall, UTR-LM demonstrates superior performance that is robust and stable across random repeats. Note the MRL value ranges in [0,12] in the dataset. For the UTR-LM, data are presented as mean values +/- SD (n=200,000).

E. Model Interpretation

E.1. UTR-LM embeddings for differentiating between species

We tested if the pre-trained embedding can differentiate between species. For the baseline, we used the k -mer representation that is common in RNA-seq analysis⁴⁹ due to its ability to capture critical features such as secondary structure, splice sites, and binding sites of RNA-binding proteins. We set $k=4$ to strike a balance between computation complexity and information captured. We computed silhouette scores between different species using the pretrained embedding and compared to the score computed using the 4-mers. Fig. 6(b) showed that the UTR-LM embedding can better capture the dissimilarity between species.

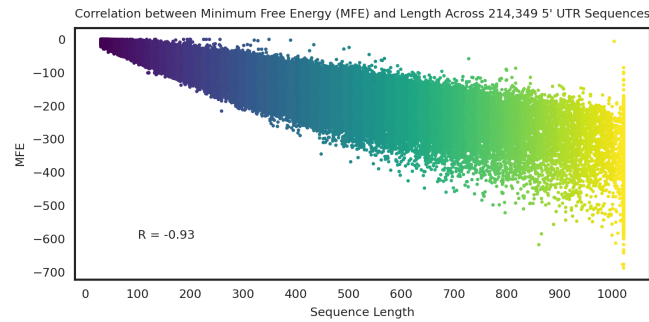
Further, we visualized UTR-LM's embeddings of 5' UTR sequences across species in Supplementary Fig. 6. We observe that the UTR embeddings can visually distinguish zebrafish from other species. It also tells noticeable differences between humans and other species. The contrast between chicken, rat and mouse are less visually obvious. These observations are consistent with the quantitative contrast between species measured via silhouette scores; see Fig. 6(b).



Supplementary Fig. 6. The visualization of the UTR-LM representation of 5' UTRs of five species.

E.2. Correlation between MFE and sequence length

The 5' UTR length is correlated with its minimum fold energy (MFE), as shown in Supplementary Fig. 7. This finding aligns with the established understanding that longer sequences are on average more stable because they can form more stacking and hydrogen bond interactions.



Supplementary Fig. 7. The correlation between Minimum Free Energy (MFE) and sequence length across 214,349 5' UTR sequences from five species.

E.3. Attention Score Analysis and Biological Findings

We examined the attention scores learnt in the UTR-LM model. In particular, we illustrate two examples: one 5' UTR with high MRL value and another one with low MRL value, to better demonstrate the model's ability to score important sites. For comparison, we also provided untrained attention scores for these two examples. See Supplementary Fig. 8 for these examples.

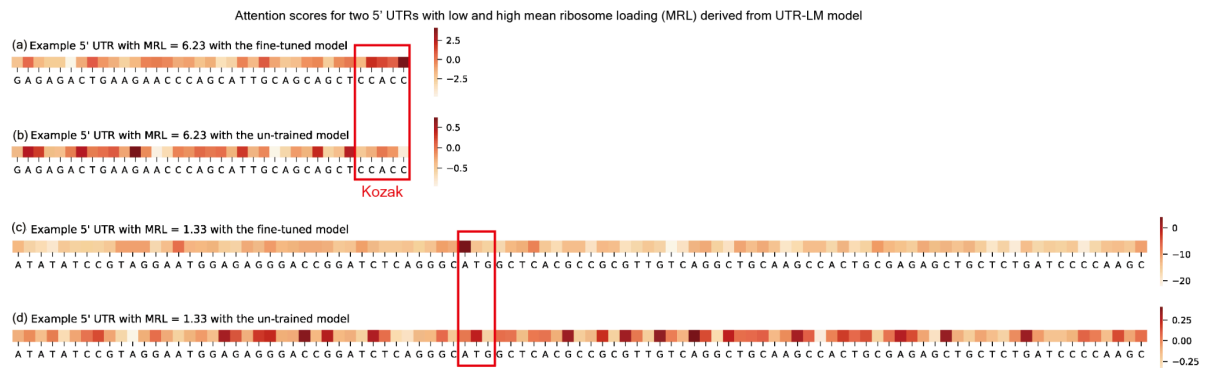
Supplementary Fig. 8 (a,b) gave a UTR sequence with a high MRL value of 6.23. The attention score of our model highlighted the crucial Kozak motif (CCACC) at the rightmost, which is well known as a key motif for initiating translation. These examples demonstrate the fine-tuned UTR-LM model's ability to recognize regulatory motifs impacting translation efficiency.

Supplementary Fig. 8(c,d) showed an UTR sequence with a low MRL value 1.33. The attention scores highlighted an ATG codon in the middle of the sequence. The ATG, possibly hindering normal coding sequence translation due to serving as a competing upstream start codon (uATG) driving potential upstream open reading frame (uORF), is recognized by the fine-tuned model. Prior study has identified the importance of uORF may cause lower protein expression^{20,50}, consistent with the low MRL value of this example.

Our analysis pointed out that the positioning of the ATG within UTR may also be crucial, as the attention score from our UTR-LM pointing towards that uATG located in the middle of upstream UTR region (positions 58-60) may have a stronger effect over downstream protein expression. While further studies are needed to validate this, we believe this showcases the ability of sequence language models to reveal new functional insights that could be biologically interpretable and relevant⁵¹.

Further, we found a significant difference in ATG occurrences between the lowest and highest 10% of mean ribosome loading (MRL) sequences (n=1,555). The lowest 10% had 1,363 ATG occurrences, averaging 1.52 per sequence, compared to only 95 occurrences, averaging 1.08 per sequence, in the highest 10%. This significant difference was confirmed by a paired t-test (p -value < 0.05). Supplementary Fig. 9 (b,c) shows a contrast in 5' UTRs between the lowest 10% MRLs with more ATG occurrences and the highest 100 MRLs with fewer.

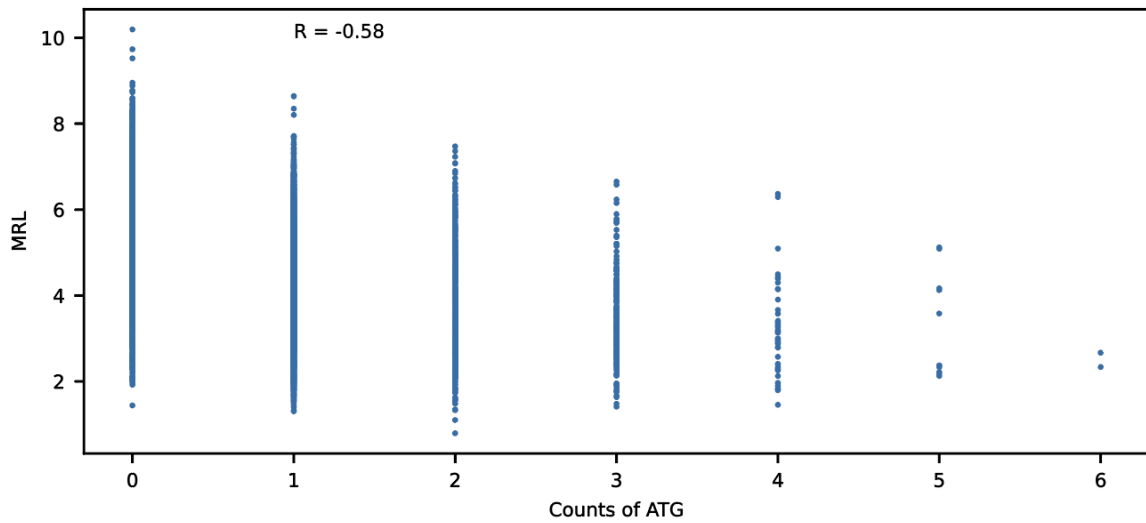
Interestingly, in low-MRL sequences, ATG positions were more frequently located on the left (far from the CDS), while in high-MRL sequences, they were more commonly found in the middle.



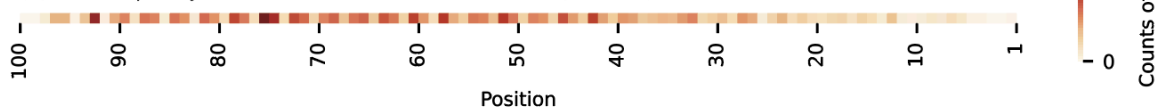
Supplementary Fig. 8. Examples of two 5' UTRs with low and high values of mean ribosome loading (MRL), respectively, and their attention scores learnt in the UTR-LM model, respectively. (a,b) A high MRL sequence, where the attention scores highlight the rightmost Kozak motif (CCACC) that is known to play a critical role in translation. (c,d) A sequence with low MRL value, where the trained attention scores highlight the central ATG codon that is known to hinder the translation. These important motifs are successfully identified in the fine-tuned model (a,c) but are not detected in the untrained model (b,d).

Correlation between ATG occurrences and Mean Ribosome Loading (MRL) across 15,555 Human 5' UTR Sequences

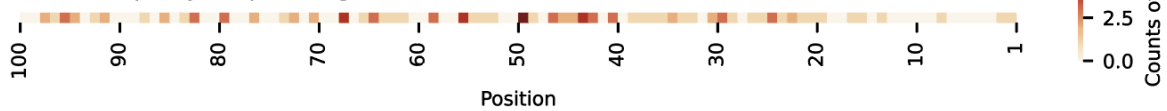
(a) Correlation between ATG occurrences and MRL



(b) ATG Frequency in Bottom 10% Lowest MRL 5' UTRs (n=1555)



(c) ATG Frequency in Top 10% Highest MRL 5' UTRs (n=1555)

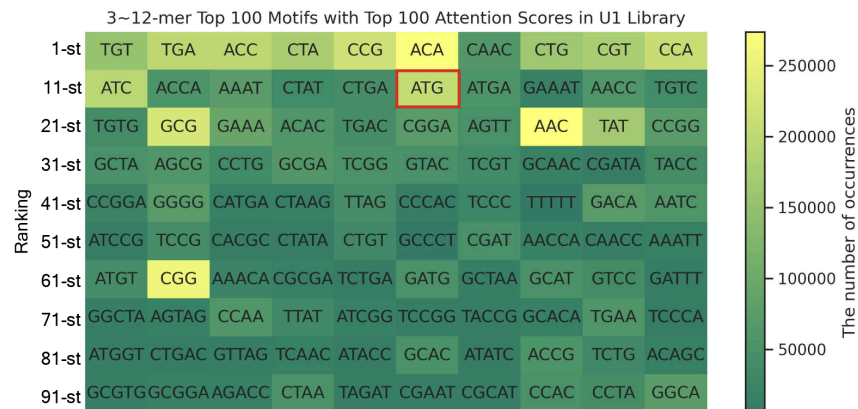


Supplementary Fig. 9. Correlation between ATG occurrences and mean ribosome loading (MRL) across 15,555 human 5' UTR sequences. (a) It reveals a pattern where higher ATG occurrences generally correspond to lower MRL. (b) In 5' UTRs with the bottom 10% lowest MRLs, a higher frequency of ATG occurrences was observed, predominantly on the left side, farther from the CDS. (c) Conversely, in 5' UTRs with the top 10% highest MRLs, which had fewer ATG occurrences, these were more commonly located in the middle.

E.4. Exhaustive scan of 3-12 mer fragments

We ran a screen of all 3-12 mer fragments, filtered and ranked by attention scores, to identify new potential motifs associated with Mean Ribosome Loading (MRL). To achieve this, we introduced a novel convolution operation to remove the positional effect of the attention mechanism, hoping to detect global motifs independent of their position. The technique involved calculating the cumulative attention score for each k -mer motif throughout the sequences, adjusting for motif length. By doing so, we obtained a list of k -mer fragments as potential motifs, ranked by their attention scores and occurrences. See Supplementary Fig. 10 for top 100 candidates of such motifs, and please find the full list [here](#). A higher accumulative attention score suggests a stronger relationship between the candidate motif and the

predicted MRL. While these motif patterns call for additional validation and biological studies, they hold the potential to unlock new understanding into mRNA translation.



Supplementary Fig. 10. Top 100 motif patterns ranked by attention scores associated with MRL prediction, where the color intensity denotes the motif's frequency of occurrence. The highlighted ATG motif aligns with our finding in E.3. The complete list can be found [here](#).

References

1. Araujo, P. R. *et al.* Before It Gets Started: Regulating Translation at the 5' UTR. *Comp. Funct. Genomics* **2012**, 475731 (2012).
2. Miao, Z., Tidu, A., Eriani, G. & Martin, F. Secondary structure of the SARS-CoV-2 5'-UTR. *RNA Biol.* **18**, 447–456 (2021).
3. Li, X., Kazan, H., Lipshitz, H. D. & Morris, Q. D. Finding the target sites of RNA-binding proteins. *Wiley Interdiscip. Rev. RNA* **5**, 111–130 (2014).
4. Zeraati, M. *et al.* Cancer-associated noncoding mutations affect RNA G-quadruplex-mediated regulation of gene expression. *Sci. Rep.* **7**, 708 (2017).
5. Karollus, A., Avsec, Ž. & Gagneur, J. Predicting mean ribosome load for 5'UTR of any length using deep learning. *PLoS Comput. Biol.* **17**, e1008982 (2021).
6. Sample, P. J. *et al.* Human 5' UTR design and variant effect prediction from a massively parallel translation assay. *Nat. Biotechnol.* **37**, 803–809 (2019).
7. Cao, J. *et al.* High-throughput 5' UTR engineering for enhanced protein production in non-viral gene therapies. *Nat. Commun.* **12**, 4138 (2021).
8. Barazandeh, S., Ozden, F., Hincer, A., Seker, U. O. S. & Cicek, A. E. UTRGAN: learning to generate 5' UTR sequences for optimized translation efficiency and gene expression. *BioRxiv* (2023) doi:10.1101/2023.01.30.526198.
9. Zheng, W. *et al.* Discovery of regulatory motifs in 5' untranslated regions using interpretable multi-task learning models. *Cell Syst.* (2023) doi:10.1016/j.cels.2023.10.011.
10. Chen, J. *et al.* Interpretable RNA Foundation Model from Unannotated Data for Highly Accurate RNA Structure and Function Predictions. *arXiv* (2022) doi:10.48550/arxiv.2204.00300.
11. Ozden, F., Barazandeh, S., Akboga, D., Seker, U. O. S. & Cicek, A. E. RNAGEN: A generative adversarial network-based model to generate synthetic RNA sequences to target proteins. *BioRxiv* (2023) doi:10.1101/2023.07.11.548246.
12. Akiyama, M. & Sakakibara, Y. Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. *NAR Genom. Bioinform.* **4**, lqac012 (2022).
13. Wang, J. & Gribskov, M. IRESpy: an XGBoost model for prediction of internal ribosome entry sites. *BMC Bioinformatics* **20**, 409 (2019).
14. Kolekar, P., Pataskar, A., Kulkarni-Kale, U., Pal, J. & Kulkarni, A. Irespred: web server for prediction of cellular and viral internal ribosome entry site (IRES). *Sci. Rep.* **6**, 27436 (2016).
15. Zhao, J. *et al.* IRESfinder: Identifying RNA internal ribosome entry site in eukaryotic cell using framed k-mer features. *J. Genet. Genomics* **45**, 403–406 (2018).
16. Zhou, Y. *et al.* DeepCIP: a multimodal deep learning method for the prediction of internal ribosome entry sites of circRNAs. *BioRxiv* (2022) doi:10.1101/2022.10.03.510726.
17. Zeng, C. *et al.* Leveraging mRNA Sequences and Nanoparticles to Deliver SARS-CoV-2 Antigens In Vivo. *Adv. Mater.* **32**, e2004452 (2020).

18. Babendure, J. R., Babendure, J. L., Ding, J.-H. & Tsien, R. Y. Control of mammalian translation by mRNA structure near caps. *RNA* **12**, 851–861 (2006).
19. Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science* **352**, 1413–1416 (2016).
20. Calvo, S. E., Pagliarini, D. J. & Mootha, V. K. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci USA* **106**, 7507–7512 (2009).
21. *Post-Transcriptional Gene Regulation*. vol. 1358 (Springer New York, 2016).
22. Whiffin, N. *et al.* Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. *BioRxiv* (2019) doi:10.1101/543504.
23. Kozak, M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* **15**, 8125–8148 (1987).
24. Kozak, M. Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. *Proc Natl Acad Sci USA* **87**, 8301–8305 (1990).
25. Stoneley, M. & Willis, A. E. Cellular internal ribosome entry segments: structures, trans-acting factors and regulation of gene expression. *Oncogene* **23**, 3200–3207 (2004).
26. Weingarten-Gabbay, S. *et al.* Comparative genetics. Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science* **351**, (2016).
27. Zhao, J. *et al.* Iresbase: A comprehensive database of experimentally validated internal ribosome entry sites. *Genomics Proteomics Bioinformatics* **18**, 129–139 (2020).
28. Mokrejs, M. *et al.* IRESite--a tool for the examination of viral and cellular internal ribosome entry sites. *Nucleic Acids Res.* **38**, D131-6 (2010).
29. Kalvari, I. *et al.* Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **49**, D192–D200 (2021).
30. Leppek, K. *et al.* Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. *Nat. Commun.* **13**, 1536 (2022).
31. Gleason, A. C., Ghadge, G., Chen, J., Sonobe, Y. & Roos, R. P. Machine learning predicts translation initiation sites in neurologic diseases with nucleotide repeat expansions. *PLoS ONE* **17**, e0256411 (2022).
32. Hernández, G., Osnaya, V. G. & Pérez-Martínez, X. Conservation and variability of the AUG initiation codon context in eukaryotes. *Trends Biochem. Sci.* **44**, 1009–1021 (2019).
33. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).
34. Vaswani, A. *et al.* Attention is All you Need. *Advances in Neural Information Processing Systems* (2017).
35. Sinha, K. *et al.* Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little. *arXiv* (2021) doi:10.48550/arxiv.2104.06644.
36. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
37. Leppek, K., Das, R. & Barna, M. Functional 5' UTR mRNA structures in

- eukaryotic translation regulation and how to find them. *Nat. Rev. Mol. Cell Biol.* **19**, 158–174 (2018).
38. Rao, R. M., Meier, J., Sercu, T., Ovchinnikov, S. & Rives, A. Transformer protein language models are unsupervised structure learners. *BioRxiv* (2020) doi:10.1101/2020.12.15.422761.
 39. Iuchi, H. *et al.* Representation learning applications in biological sequence analysis. *Comput. Struct. Biotechnol. J.* **19**, 3198–3208 (2021).
 40. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA* **118**, (2021).
 41. Rao, R. *et al.* Evaluating Protein Transfer Learning with TAPE. *Adv. Neural Inf. Process. Syst.* **32**, 9689–9701 (2019).
 42. Lin, Z. *et al.* Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv* (2022) doi:10.1101/2022.07.20.500902.
 43. Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
 44. Benegas, G., Batra, S. S. & Song, Y. S. DNA language models are powerful zero-shot predictors of non-coding variant effects. *BioRxiv* (2022) doi:10.1101/2022.08.22.504706.
 45. Dalla-Torre, H. *et al.* The nucleotide transformer: building and evaluating robust foundation models for human genomics. *BioRxiv* (2023) doi:10.1101/2023.01.11.523679.
 46. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).
 47. Chen, K. *et al.* Self-supervised learning on millions of pre-mRNA sequences improves sequence-based RNA splicing prediction. *BioRxiv* (2023) doi:10.1101/2023.01.31.526427.
 48. Wang, X., Gao, X., Wang, G. & Li, D. miProBERT: identification of microRNA promoters based on the pre-trained model BERT. *Brief. Bioinformatics* (2023) doi:10.1093/bib/bbad093.
 49. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
 50. Jones, L. *et al.* Translational effects and coding potential of an upstream open reading frame associated with DOPA Responsive Dystonia. *Biochim. Biophys. Acta Mol. Basis Dis.* **1863**, 1171–1182 (2017).
 51. Schaeffer, R., Miranda, B. & Koyejo, S. Are Emergent Abilities of Large Language Models a Mirage? *arXiv* (2023) doi:10.48550/arxiv.2304.15004.