Acta Pharmaceutica Sinica B 2024;14(4):1814–1826

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com

ORIGINAL ARTICLE

A novel deep generative model for mRNA vaccine development: Designing 5′ UTRs with *N*1-methyl-pseudouridine modification

Xiaoshan Tang[a,†], Miaozhe Huo[b,†], Yuting Chen[a,†], Hai Huang[a,†], Shugang Qin[a], Jiaqi Luo[b], Zeyi Qin[c], Xin Jiang[a], Yongmei Liu[a], Xing Duan[a], Ruohan Wang[b], Lingxi Chen[b], Hao Li[a], Na Fan[a], Zhongshan He[a], Xi He[a], Bairong Shen[a,*], Shuai Cheng Li[b,*], Xiangrong Song[a,*]

[a]*Institute of Systems Genetics, Department of Critical Care Medicine, Frontiers Science Center for Disease-related Molecular Network, State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, Sichuan University, Chengdu 610000, China*
[b]*Department of Computer Science, City University of Hong Kong, Hong Kong 99907, China*
[c]*Department of Biology, Brandeis University, Boston, MA 02453, USA*

Received 24 August 2023; received in revised form 26 October 2023; accepted 1 November 2023

KEY WORDS

mRNA vaccine;
Machine learning;
5′ UTR;
mRNA design;
Sequence design;
*N*1-Methyl-pseudouridine;
COVID-19;
SARS-CoV-2

Abstract Efficient translation mediated by the 5′ untranslated region (5′ UTR) is essential for the robust efficacy of mRNA vaccines. However, the *N*1-methyl-pseudouridine (m1Ψ) modification of mRNA can impact the translation efficiency of the 5′ UTR. We discovered that the optimal 5′ UTR for m1Ψ-modified mRNA (m1Ψ−5′ UTR) differs significantly from its unmodified counterpart, highlighting the need for a specialized tool for designing m1Ψ−5′ UTRs rather than directly utilizing high-expression endogenous gene 5′ UTRs. In response, we developed a novel machine learning-based tool, Smart5UTR, which employs a deep generative model to identify superior m1Ψ−5′ UTRs *in silico*. The tailored loss function and network architecture enable Smart5UTR to overcome limitations inherent in existing models. As a result, Smart5UTR can successfully design superior 5′ UTRs, greatly benefiting mRNA vaccine development. Notably, Smart5UTR-designed superior 5′ UTRs significantly enhanced antibody titers induced by COVID-19 mRNA vaccines against the Delta and Omicron variants of SARS-CoV-2, surpassing the performance of vaccines using high-expression endogenous gene 5′ UTRs.

*Corresponding authors.
E-mail addresses: bairong.shen@scu.edu.cn (Bairong Shen), shuaicli@cityu.edu.hk (Shuai Cheng Li), songxr@scu.edu.cn (Xiangrong Song).
[†]These authors made equal contributions to this work.

Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2023.11.003

1. Introduction

In recent decades, pandemics of infectious diseases have constantly occurred, including severe acute respiratory syndrome in 2002, influenza in 2009, Middle East respiratory syndrome in 2012, Ebola in 2014, Zika in 2015, and COVID-19 in 2019[1,2]. During the COVID-19 pandemic, mRNA technology has accelerated the development and production of vaccines at unprecedented speed, which protects billions of people in a timely and safe manner (www.who.int). Although it is evident that mRNA vaccines represent a unique and versatile platform for combating infectious diseases, further research is needed to improve the design of mRNA vaccine design[3,4].

Sufficient protein expression is the prerequisite of mRNA vaccines to induce strong immune responses[4,5]. The translation of mRNA was predominantly impacted by the 5′ untranslated region (5′ UTR). The superior 5′ UTR sequence can guarantee the expression of mRNA and the efficacy of the mRNA vaccine[6,7]. The mechanism by which 5′ UTR sequences regulate mRNA translation remains unclear, hindering *de novo* design[8]. Traditionally, researchers used 5′ UTRs of endogenous genes with high expression levels (endogenous superior 5′ UTR) in mRNA medicine[9−11]. However, the *N*1-methyl-pseudouridine (m1Ψ) modification influences mRNA translation[12,13], suggesting that the superior sequence of endogenous 5′ UTRs may differ from that of m1Ψ-modified mRNA. Notably, m1Ψ-modified mRNA plays a crucial role in vaccine performance[14] and is a cornerstone technology in mRNA medicine. Therefore, identifying the superior 5′ UTR sequence for m1Ψ-modified mRNA (m1$\Psi-5'$ UTR) is essential. Inherently, the m1$\Psi-5'$ UTR design is to find optimal nucleotide combinations in 4^N possible combinations (where N is the nucleotide sequence length), which is a combinational optimization problem. With such a large number of potential base combinations, deep learning offers a practical solution for searching robust m1$\Psi-5'$ UTR[15−17]. Sample et al.[18] reported a generative model for 5′ UTR, termed Optimus 5-Prime, which used the convolutional neural network (CNN) and the genetic algorithm (GA) to search for superior unmodified 5′ UTR. However, Optimus 5-Prime is inefficient, in which data fitting and generative modules are separated, and search results largely depend on the random initial population.

In this study, we showed the difference between unmodified and m1Ψ-modified 5′ UTRs by implementing a detailed analysis of the library reported by Sample and colleagues[18]. Our findings revealed that m1Ψ-modified 5′ UTRs exhibited distinct patterns, necessitating the development of models tailored to m1Ψ-modified 5′ UTRs. We developed an all-in-one model, Smart5UTR, to search for superior m1$\Psi-5'$ UTR sequences *in silico*. Smart5UTR can efficiently generate the 5′ UTR sequences based on the multitask autoencoder (MTAE) frame, which can fully exploit the features learned from input 5′ UTRs. To validate Smart5UTR *in vitro*, we synthesized a series of m1$\Psi-$mRNAs containing 5′ UTRs with different translational efficiency, and they showed relative expression levels. Furthermore, we engineered COVID-19 mRNA vaccines that incorporate superior 5′ UTRs designed by Smart5UTR and the endogenous superior 5′ UTR (S27a-44′ 5′ UTR) screened out by Zeng and colleagues[9]. The Smart5UTR-designed vaccines induced a stronger immune response against the Delta and Omicron variants of SARS-CoV-2 compared to the one with endogenous superior 5′ UTR. In conclusion, we developed a novel tool for m1$\Psi-5'$ UTR design and performed proof-of-concept experiments based on COVID-19 mRNA vaccines (Fig. 1).

2. Materials and methods

2.1. *Development of the MTAE architecture for 5′ UTR sequence reconstruction and MRL prediction*

An MTAE model was built to interactively predict the MRL value for a 50 nt 5′ UTR sequence and a novel 5′ UTR sequence *de novo* design. Given an input 5′ UTR sequence X_{UTR} of length (L), the one-hot encoding method generates a two-dimensional $L \times 4$ matrix X_{onehot} as in Eq. (1):

$$X_{\text{onehot}}(i,j) = \begin{cases} 1, \text{if riboside } R_i \text{ corresponds to the } j\text{−th riboside} \\ 0, \text{otherwise} \end{cases} \tag{1}$$

Here, we encode A (adenine riboside) as (1 0 0 0), C (cytosine riboside) as (0 1 0 0), G (guanine riboside) as (0 0 1 0), and U/m1Ψ (uracil riboside/*N*1-methyl-pseudouracil riboside) as (0 0 0 1).

The encoder contains four 1-D convolutional layers with 160, 160, 160, and 80 kernels of length eight, followed by batch normalization layers. Next, two fully connected layers with dropout are applied, and a linear activation function is used for the final output. The resulting 80-dimensional latent vector Z as in Eq. (2) represents the input UTR sequence:

$$Z = \text{Encoder}(X_{\text{onehot}}) \tag{2}$$

A decoder is added to the regression model, which accepts the concatenation of the latent vector Z and twenty repeated output nodes y_{out} from the regressor as input. Denote the concatenation result as vector C as in Eq. (3):

$$C = \text{Concatenate}(Z, y_{\text{out}}) \tag{3}$$

The decoder includes fully connected 1D convolutional layers, ultimately reproducing a matrix Y_{pred} with shape $L \times 4$ as output. The decoder function can be represented as in Eq. (4):

$$X_{\text{pred}} = \text{Decoder}(C) \tag{4}$$

The model is trained on the EGFP1 MPRA dataset from Sample et al[18]. The encoder aims to fit the MRL labels using an adjusted weighted mean-squared error loss function L_{WMSE}, while the decoder minimizes the reconstruction error using a categorical cross-entropy loss function L_{CE}. The total loss function is a linear combination of these two loss terms, formulated as in Eq. (5):

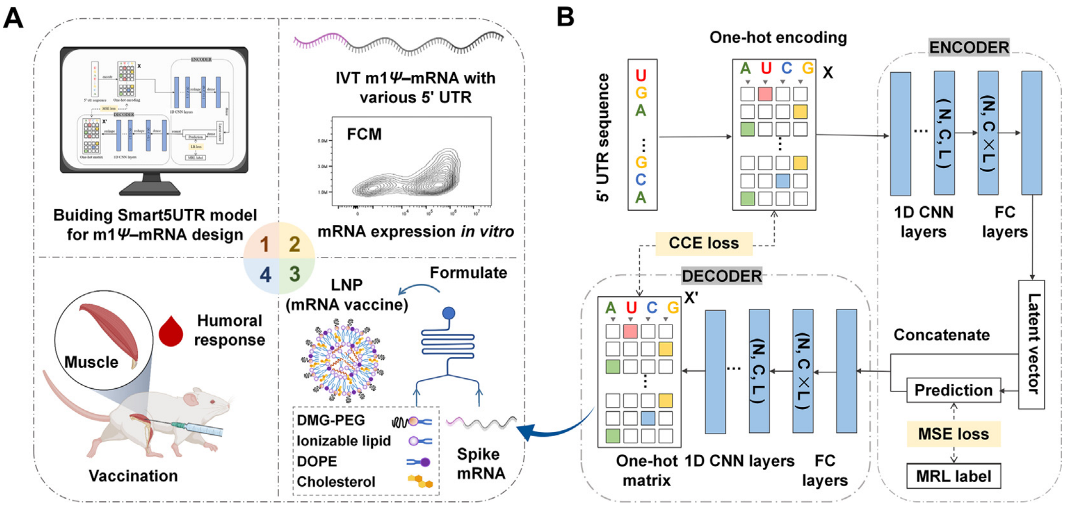


Figure 1 Overview of Smart5UTR for the design of mRNA vaccines. (A) The process of building and evaluating a multi-task autoencoder model, Smart5UTR. Smart5UTR was built for mRNA design and validated *in vitro* using flow cytometry to detect the protein expression of mRNAs with various 5′ UTRs. Then, Smart5UTR engineered the mRNA encoding the spike protein for *in vivo* evaluation. Spike mRNAs were synthesized and formulated into LNPs in microfluidic devices. Mice were intramuscularly vaccinated with mRNA vaccines, and serum was collected to evaluate the humoral response. (B) The architecture of Smart5UTR. Each 50 nt 5′ UTR sequence was encoded into a matrix by one-hot encoding as the model's input. 1D CNN layers followed by FC layers mapped the input matrix to a latent vector. A linear output node maps to the corresponding MRL label and was trained with a loss of MSE. Then the latent vector concatenated with its linear output and was given as input to 1D CNN layers and FC layers to generate a novel 5′ UTR. CCE loss was used to ensure the similarity of the input and output sequences.

$$\mathrm{Loss} = \alpha \cdot L_{\mathrm{WMSE}}\left(y, y_{\mathrm{pred}}\right) + \beta \cdot L_{\mathrm{CE}}\left(X_{\mathrm{onehot}}, X_{\mathrm{pred}}\right) \tag{5}$$

Initially, the model was trained for 20 epochs with a learning rate of $2\mathrm{e}^{-4}$ and loss weights of 1:5. Training then continued for 80 epochs at a learning rate of $1\mathrm{e}^{-5}$ and equal loss weights.

2.2. *Loss functions in the MTAE model*

We performed z-score normalization on the MRL values of the whole training dataset. In this way, a raw MRL value x is converted into the standard value z according to Eq. (6):

$$z = \frac{x - \mu}{\sigma} \tag{6}$$

where μ is the mean of the MRL and σ represents the standard deviation of the MRL value.

We customized a weighted MSE loss function to train the regressor. For a data set of size n, the loss function was defined as in Eqs. (7) and (8):

$$L_{\mathrm{WMSE}} = \frac{1}{n} \sum_{i=1}^{n} \alpha \left(y_i - \widehat{y_i}\right)^2 \tag{7}$$

$$\alpha = \begin{cases} 1, & z < 0 \\ 1 + z, & z \geq 0 \end{cases} \tag{8}$$

where y_i is the true MRL value for each data i and $\widehat{y_i}$ is the predicted value from the MTAE model. This weighted loss function design allows the model to focus more on data points with larger MRL values during training, thereby improving the model's accuracy in predicting larger MRL values.

We used the categorical cross-entropy loss function to minimize the reconstruction error between the input and reconstructed UTR sequences, which is computed as in Eq. (9):

$$L_{\mathrm{CE}} = -\sum_{i=1}^{n} T_i \log(S_i) \tag{9}$$

where $n = 4$, representing the total number of nucleotide classes, T_i represents the true label of the i-th nucleotide class and S_i is the predictive probability of this class.

2.3. *Reimplement the benchmark models*

Two other machine learning algorithms were used to computationally capture the relationship between the m1Ψ−5′ UTR sequences and the MRL values, including the CNN frame and the random forest models. Both models were trained on the same MPRA dataset as Smart5UTR. We utilized the top 200,000 sequences with the highest total reads to train and evaluate models (the UTRs with higher read counts reflect their MRL more accurately). The CNN-frame predictor was reimplemented referencing the work of Sample et al.[18], named Optimus 5-Prime. The random forest model (referring to Cao's[19]) took the 3-mer frequency, the RNA folding energy, and the number of upstream Open reading frames (uORFs) extracted from 50 nt synthetic as the features.

2.4. *Optimization of 5′ UTR based on MTAE frame*

We sampled 50 nt prototypical 5′ UTR sequences from the test dataset and optimized them based on the Smart5UTR framework. The one-hot encoded prototypes were fed into the Smart5UTR to generate latent vectors first. Let x_i denote the one-hot encoded input sequence, and Encoder represents the encoder function of the Smart5UTR, the latent vector for the i-th sequence, v_i, can be obtained as in Eq. (10):

$$v_i = \mathrm{Encoder}(x_i) \tag{10}$$

Next, we introduce a coefficient, γ, to control the degree of MRL enhancement when optimizing the input sequence. To achieve this, we adjust the original latent vector v_i by multiplying each element with a factor that includes a Gaussian noise term modulated by parameters α and γ. Furthermore, we concatenate a vector containing γ repeated n times to increase the influence of γ in the generation process. The modified latent vector v_i' is given as in Eq. (11):

$$v_i' = v_i \cdot \left(\alpha \cdot \gamma + \mathcal{N}\left(0, \sigma^2\right)\right) \oplus \underbrace{(\gamma, \ldots, \gamma)}_{\text{repeat } n} \tag{11}$$

where α is a parameter to control the degree of Gaussian noise and $\mathcal{N}(0, \sigma^2)$ represents the normal distribution with mean 0 and variance σ^2.

These modified latent vectors v_i' are then passed through the decoder, denoted by Decoder, to generate novel candidate UTR sequences by

$$\widehat{x_i} = \text{Decoder}\left(v_i'\right) \tag{12}$$

We select generated sequences with higher than those of their corresponding prototypes, resulting in a new set of sequences. This optimization step is repeated for a fixed number of iterations or until there are only minuscule changes in the set of sequences.

2.5. *Optimization of 5′ UTR based on genetic algorithm*

We applied a genetic algorithm (GA) adapted from Sample et al.[18] to search for optimal 50 nt 5′ UTR sequences. To ensure a fair comparison, we made minor adjustments to the original GA while retaining its core principles. Denote x_i as the i-th input UTR sequence with MRL y_i, and $M(x_i,p)$ represent the mutation function that modifies x_i with a mutation probability p. The GA implementation mutates nucleotides in the input UTR to generate a candidate sequence x_i' as in Eq. (13):

$$x_i' = M(x_i, p) \tag{13}$$

where we have modified the original default mutation probability of 50% to include two options, 50% and 70%. This change allows for greater flexibility in exploring the search space of 5′ UTR sequences.

After mutation, if the predicted MRL $(\widehat{y_i'})$ of x_i' is more significant than y_i, the new sequence is considered for the next iteration (evolution), as in Eq. (14):

$$x_i^{(t+1)} = \begin{cases} x_i', \widehat{y_i'} > y_i \\ x_i, \quad \text{otherwise} \end{cases} \tag{14}$$

where let $x_i^{(t+1)}$ represent the updated value of the i-th input at time step $t+1$, $\widehat{y_i'}$ be the predicted value of the i-th input based on the updated feature vector x', and y_i denote the actual value of the i-th input feature.

The adapted GA starts with an initial set of sequences and iteratively mutates and evaluates them using Optimus 5-Prime. If the candidate sequence has a higher MRL value than the original sequence, it is accepted and incorporated into the new set for subsequent iterations. The GA iteratively optimizes 5′ UTRs until convergence is achieved or the maximum number of iterations is reached. By making these adjustments, we aim to provide a more balanced comparison between our proposed method and the one described in the Optimus 5-Prime literature.

2.6. *In vitro mRNA transcription*

All mRNAs were synthesized using T7 RNA polymerase (Vazyme, DD4101, Nanjing, China), linearizing plasmid DNA templates, nucleoside triphosphates (*N*1-methyl-pseudouridine replacing uridine, TriLink, N-1081, San Diego, CA, USA), and a cap1 analog (TriLink, CleanCap® AG (3′ OMe), San Diego, CA, USA). After digesting DNA templates with DNase (Vazyme, D4104, Nanjing, China), mRNAs were extracted using an RNA extracting solution (Solarbio, P1011, Beijing, China) and precipitated in 2.5 mol/L ammonium acetate solution. The same 3′ UTR sequence was used in all mRNAs with different 5′ UTRs (Supporting Information Table S1). The mRNAs encoding EGFP included a 50 nt poly(A) tail and the ones encoding the spike protein from SARS-CoV-2 (B.1.617.2) had a 100 nt poly(A) tail. The mRNA encoding spike protein was designed in reference to the SARS-CoV-2 B.1.617.2 variant (GenBank OK091006).

2.7. *In vitro expression of mRNA*

For transfection of mRNA into HEK293T cells (ATCC, CRL-3216™) or DC2.4 cells, 1 × 10^5 cells per well were seeded in 24-well plates and cultured with 0.5 mL of DMEM (Gibco, A4192101, Grand Island, NY, USA) with 10% (*v*/*v*) FBS (Gibco, Grand Island, 10091148, NY, USA) and 1% (*v*/*v*) antibiotics (Hyclone, SV30010, Logan, UT, USA) in overnight. One microgram of EGFP-mRNAs with various 5′ UTRs was mixed with Lipofectamine2000 (Thermo Fisher Scientific, 11668027, Waltham, MA, USA) and added to the corresponding well after 30-min incubation. After 24 h, we performed flow cytometry to evaluate transfection efficacy. DC2.4 cells were provided by Prof. Zhen Gu (Zhejiang University, Zhejiang, China).

2.8. *LNP preparation and characterization*

Lipid nanoparticles (LNPs) were prepared by mixing the mRNA aqueous phase and lipid ethanol phase at a volume ratio of 3:1 with the microfluidic equipment (Precision NanoSystems, Vancouver, Canada) at the flow rate of 9 mL/min. The *N*/*P* ratio was 17:1. The mRNA dissolved in 10 mmol/L citrate buffer (pH = 3.0). The control LNP was mRNA free using 10 mmol/L citrate buffer (pH = 3) as aqueous phase. The lipid ethanol phase contains DMG-PEG2000 (AVT, O02005, Shanghai, China), DOPE (AVT, S03005, Shanghai, China), cholesterol (AVT, O01001, Shanghai, China) and ionizable lipid in the molecule ratio of 2.5:16:46.5:35. The synthesis of ionizable lipid was entrusted to HitGen Inc. (Chengdu, China) and ^{1}H NMR of ionizable lipid was showed in Supporting Information Fig. S8A (WO2022218295). The LNPs were dialyzed using 10 mmol/L citric acid buffer (pH = 6) with 6% sucrose (*w*/*w*). The measurement and calculation of the encapsulation efficiency for all formulations followed the procedure previously described[20].

2.9. *Mice immunization*

The male BALB/c mice (7–8 weeks old) were administrated with mRNA LNPs or PBS on Days 0 and 14, and serum was collected on Day 28 for analysis with enzyme-linked immunosorbent assay (ELISA). The body weight of mice was measured. The animal experiments have been approved by the Institutional Animal Care and Use Committee of West China Hospital, Sichuan University (Ethics Number: 20220330004).

2.10. Enzyme-linked immunosorbent assays

ELISA plates were coated with 100 μL of RBD protein (1 μg/mL) solution overnight at 4 °C and were blocked by 100 μL of 2% BSA in washing buffer for 4 h at 25 °C. The mice serum was performed with a 2-fold dilution beginning with a 1:100 ratio and serum dilution was added for incubating overnight at 4 °C. The anti-mice IgG antibody (Cell Signaling Technology, 7076S, Danvers, USA) was diluted at 1:50,000 and 100 μL of IgG antibody dilution was then added to ELISA plates for 2 h at 25 °C. The ELISA plates were added with 100 μL of TMB solution (Solarbio, PR1200, Beijing, China) for 0.5 h at 25 °C and 100 μL of 2 mol/L H_2SO_4 solution was used to stop the reaction. ELISA plates were read at 450 and 630 nm using a microplate reader (Tecan Group Ltd., Männedorf, Switzerland).

2.11. Quantitative enzyme-linked immunosorbent assays

After 12, 24, 48 and 72 h of transfection of vaccines in HEK293T cells and DC2.4 cells, the culture medium was collected and the S protein level was measured following the previously described procedure (Vazyme, DD3302, Nanjing, China)[21].

2.12. Quantitative PCR

After 24 h of transfection of vaccines in HEK293T cells and DC2.4 cells, the total RNA was extracted with the Total RNA kit according to the manufacturer's instructions. cDNA was obtained through reverse transcription reaction and further quantified using PowerUp SYBR Green Master Mix (Thermo Fisher Scientific, A25741, Waltham, MA, USA) and real-time quantitative PCR reaction (Bio-Rad Laboratories Ltd., Watford, Hertfordshire, UK). Spike protein: CGACGAGGTGAGACAGAT CG (forward primer), TTTCCGCCCACCTTACTGTC (reverse primer); GAPH: AGGTCGGTGTGAACGGATTTG (forward primer), GGGGTCGTTGATGGCAACA (reverse primer).

2.13. Safety evaluation

The mice's hearts, liver, spleen, lungs, and kidneys were collected two months after the second vaccination. These organs were fixed in 10% formalin (Sigma−Aldrich, PR1200, St. Louis, MO, USA). Hematoxylin/eosin staining was performed after organs were embedded in paraffin and sectioned. We collected mice serum after two months of boost vaccination for blood biochemical analysis. All biochemical indicators of blood, including creatine kinase isoenzyme MB (CKMB), alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), total protein (TP), albumin (ALB), serum creatinine (CRE), and urea (URE) were assessed by an automatic hematological biochemical analyzer (Hitachi High-Technologies Corp., Minato-ku, Tokyo, Japan).

2.14. Statistical analyses

Statistical analysis of Figs. 2 and 4−7 was performed and presented using GraphPad Prism 8.0 software, and the data of Figs. 5−7, Supporting Information Figs. S6 and S8 are shown as mean with standard deviation. Results in Fig. 6 and Supporting Information Fig. S6 are analyzed by One Way ANOVA compared with the positive control group. Statistically, significance is indicated as non-significant (ns), $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ and $^{****}P < 0.0001$. Predictive performance was evaluated using the following metrics: coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). We measured the reconstructive accuracy of Smart5UTR by the similarity between the reconstructed UTRs and the input ones.

2.15. Data availability

The scripts and models in this project are available at github. com/deepomicslab/Smart5UTR. The data used to train Smart5UTR was obtained from the public Gene Expression Omnibus database, accessible by accession number GSE114002.

3. Results

3.1. m1Ψ modification alters the 5′ UTR sequence with high translation efficiency

Multiple techniques contribute to the clinical translation of mRNA vaccines, one of which is nucleotide modification, especially the m1Ψ modification used in approved mRNA vaccines. However, chemical modification widely regulates mRNA function[22], and several recent studies found that m1Ψ modification influenced mRNA translation[12,13]. To assess the impact of m1Ψ modification, we compared the mean ribosome loading (MRL) of the same mRNA sequence with and without m1Ψ modification based on the dataset developed by Sample and colleagues[18] (Fig. 2A). The translation efficiency of mRNA was represented by MRL, which was measured through polysome profiling[18]. The m1Ψ modification had a significant effect on the translation efficacy of mRNA, and the sequences of m1Ψ−5′ UTR with high translation efficacy differed from those without the modification (Fig. 2B).

Furthermore, we analyze the influence of m1Ψ modification on the features of the 5′UTR that regulated mRNA translation. The upstream start codon (uAUG) in 5′ UTR can impair translation efficacy by competing with or sequestering the ribosome, especially when uAUG is out of frame[23,24] (Fig. 2C). We observed that the uAUG inhibited the translation efficiency of both the unmodified and m1Ψ-modified mRNA, while the inhibition of the uAUG was more significant for m1Ψ-modified mRNA (Fig. 2D and E). The decline of MRL caused by out-of-frame uAUG for m1Ψ-modified mRNA (mean MRL = 4.4997) was larger than the unmodified mRNAs (mean MRL = 4.8864). And in-frame uAUG also reduced the MRL of m1Ψ-modified mRNA (mean MRL = 5.8186) more substantially than the unmodified mRNAs (mean MRL = 6.7121). Then, we analyzed the translation initiation site (TIS) that plays an essential role in recognizing uAUG. The strong TIS enables uAUG to significantly inhibit the translation of endogenous mRNA[25]. We used 50 sequences with the lowest MRLs and 50 sequences with the highest MRLs to calculate the nucleotide frequencies of the strong and weak TIS (created by WebLogo 3[26], Supporting Information Table S2). The strongest TIS of unmodified mRNA was the same as the Kozak sequence (the strongest TIS for mammal genes), while the m1Ψ modification moderately changed the sequence of the strong TIS. The weak TIS of m1Ψ-modified mRNA also differed from the unmodified one (Fig. 2F and G).

The results showed that m1Ψ modification altered the superior mRNA sequence and the features in 5′ UTRs that influence translation, in comparison to unmodified mRNA. Therefore, it is not appropriate to design m1Ψ−5′ UTRs solely based on the information

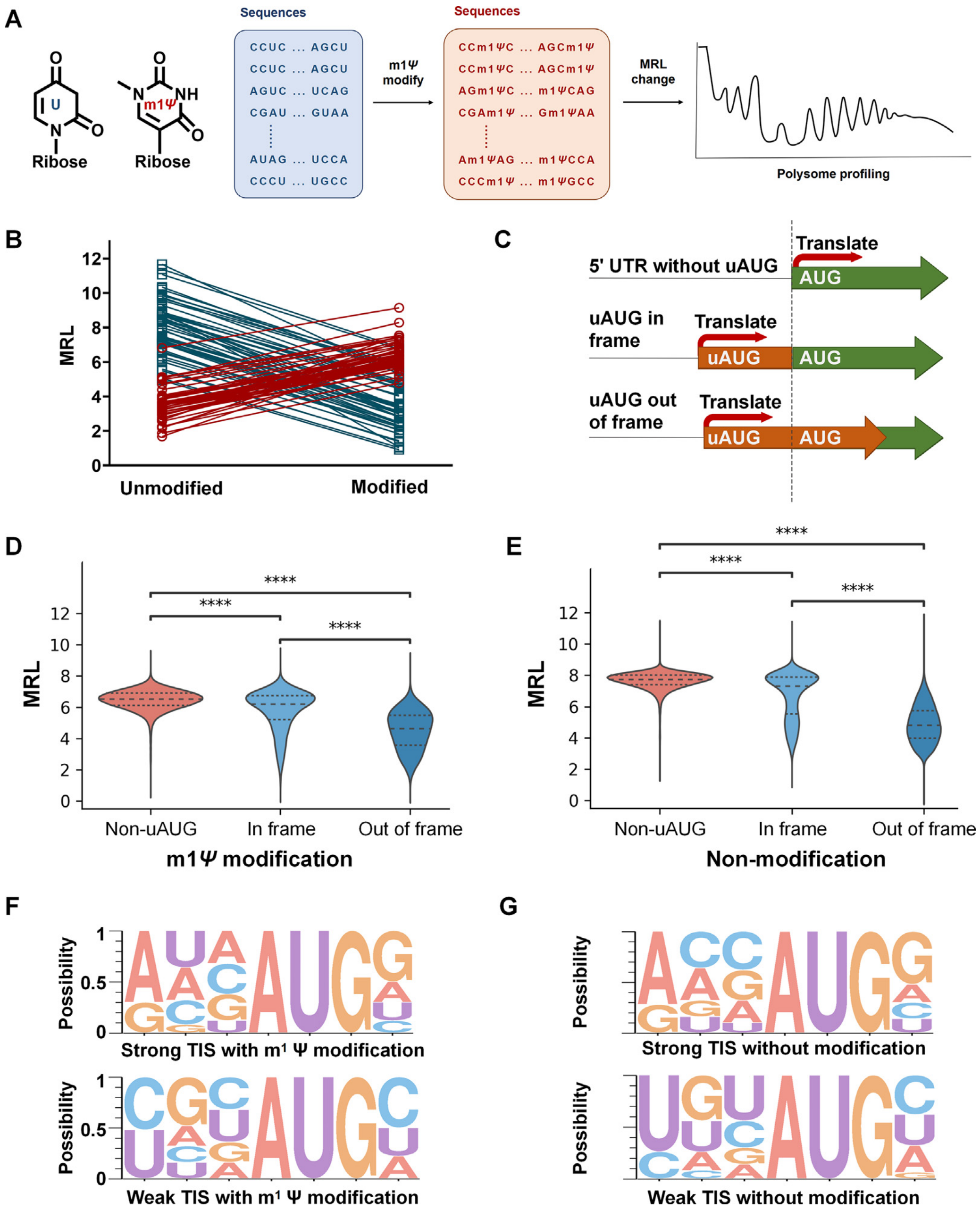


Figure 2 The influence of m1Ψ modification on 5′ UTRs. (A) Illustration of evaluating the influence of m1Ψ modification on mRNA translation. We compared two datasets that included the same mRNA sequences with m1Ψ modification and without modification. The change of translation was indicated by increased and decreased MRLs, which are assessed by polysome profiling. (B) The top 100 5′ UTRs with the most largely changed MRLs after m1Ψ modification. (C) Schematic illustration of how uAUG hinders mRNA translation. The uAUG competes with the translation of ORF (green) by producing upstream ORF (orange). (D and E) The influence of in-frame and out-of-frame uAUGs on the mean MRL of m1Ψ-modified or non-modification mRNAs. (F and G) The strong and weak TIS of m1Ψ-modified or non-modification mRNAs. WebLogo3 was used to calculate the nucleotide frequencies of the relative TIS using 50 sequences with the lowest MRL and 50 sequences with the highest MRL. $^{***}P < 0.001$, $^{****}P < 0.0001$; m1Ψ, *N*1-methyl-pseudouridine; MRL, the mean ribosome loading; uAUG, upstream initiation codon; TIS, translation initiation site.

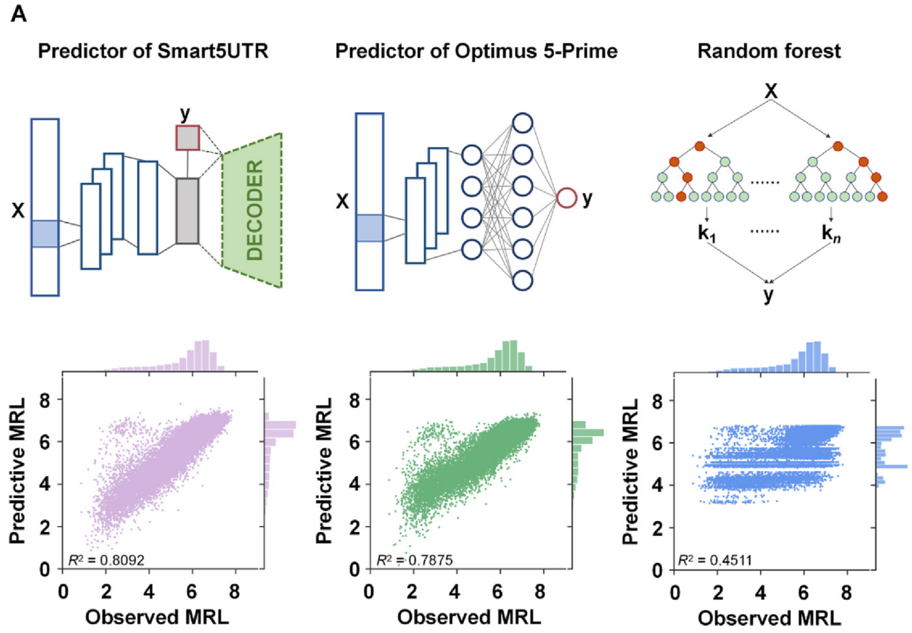
of endogenous genes without m1Ψ modification. Additional specialized tools are necessary for the m1Ψ–mRNA design.

3.2. Developing a deep learning model for m1Ψ–5′ UTR design

The design of m1Ψ–5′ UTR sequences is a combinational optimization problem where we need to find optimal solutions in a sizeable discrete space (4^N, optimize a 5′ UTR sequence with length N). Generally, models for 5′ UTR design consist of a predictor and a generator which infer the prediction of translation efficiency and the generation of new 5′ UTRs, respectively[18,19]. Nevertheless, the efficiency of existing models was limited, partially because the generator could not share the features captured by the predictor[19,27]. In comparison, Smart5UTR integrated the predictor (encoder) and the generator (decoder) that can efficiently generate the 5′ UTR sequences based on the latent vector extracted by the predictor. Moreover, we constructed a special MRL-based loss function and network architecture to render Smart5UTR a lower prediction error in the high MRL segment, which benefits the mRNA vaccine design requiring a high superior 5′ UTR sequence.

Given the importance of m1Ψ modification, we used an open dataset of m1Ψ−mRNA to develop Smart5UTR or reproduce the existing methods of deep learning (the convolutional neural network, CNN, and the random forest regression, RF). Smart5UTR, CNN and RF were trained to map 5′ UTR sequence to translation efficiency on the same datasets. Smart5UTR and CNN are CNN-based models that accept a one-hot encoding matrix for each 50 nt 5′ UTR sequence. The RF model took the k-mer frequency, the uORFs frequency and the RNA folding free energy (calculated by RNAfold[28]) as input features. Smart5UTR and Optimus 5-Prime precisely captured the relationship between m1Ψ−5′ UTRs and target MRLs (Smart5UTR, $R^2 = 0.8092$; Optimus 5-Prime, $R^2 = 0.7875$, Fig. 3A). However, RF model cannot accurately predict the MRL value ($R^2 = 0.4511$, Fig. 3A). Of the three 5′ UTR features entered into the RF model, uORFs and RNA folding energy contributed little to the MRL prediction task. uORFs are infrequent in such synthetic sequences and the correlation between RNA folding and MRL values is weak (the Pearson correlation coefficient was 0.1531 in the testing dataset, Supporting Information Fig. S1). For the latent patterns on 5′ UTR sequences, the RF model was merely trained with k-mer and the corresponding frequency, without learning the positions and interactions of the patterns, resulting in poor predictive performance.

Smart5UTR not only had the best predictive performance (Fig. 3B) and learned the relation of 5′ UTR sequence and translation (Supporting Information Fig. S2), but also had a lower prediction error in the high MRL segment, benefiting vaccine design (Fig. 4A). We achieved this improvement by constructing a special MRL-based loss function. The loss was calculated as the squared error between the predicted and actual values for each data, and for those above the average MRL value, it was multiplied by a factor that increases with the MRL value. Our customized loss function and network architecture made the regressor more sensitive to data with higher MRL values when fitting it to the input sequences. In addition to achieving the best prediction task performance, Smart5UTR successfully learned the sequence reconstruction features (5′ UTR reconstruction accuracy was 0.975). The latent vectors of Smart5UTR contained sufficient information about the ribosome binding capacity of 5′ UTRs (Fig. 4B), enabling the enhancement of protein expression of the 5′ UTR by optimizing the latent vector.



B

Metrics	Smart5UTR	Optimus 5-Prime	Random forest
R^2	0.8092	0.7875	0.4511
MAE	0.4207	0.4522	0.7009
RMSE	0.3863	0.4286	0.9755
MAPE	0.0992	0.1100	0.1721

Figure 3 The performance of models in the prediction task. (A) Smart5UTR explained 80.92% of the variability of the 5′ UTR sequences in observed MRLs. The reimplemented Optimus 5-Prime model explained 78.75% of the variability of the 5′ UTR sequences in observed MRLs. The random forest model explained 45.11% of the variability of the 5′ UTR in the observed MRLs. (B) Four metrics were calculated to compare the prediction performance of three predictors, including R^2, MAE, RMSE and MAPE. Smart5UTR had the best performance. CNN, convolutional neural network; R^2, coefficient of determination; MAE, mean absolute error; RMSE, root-mean-square error; MAPE, mean absolute percentage error.

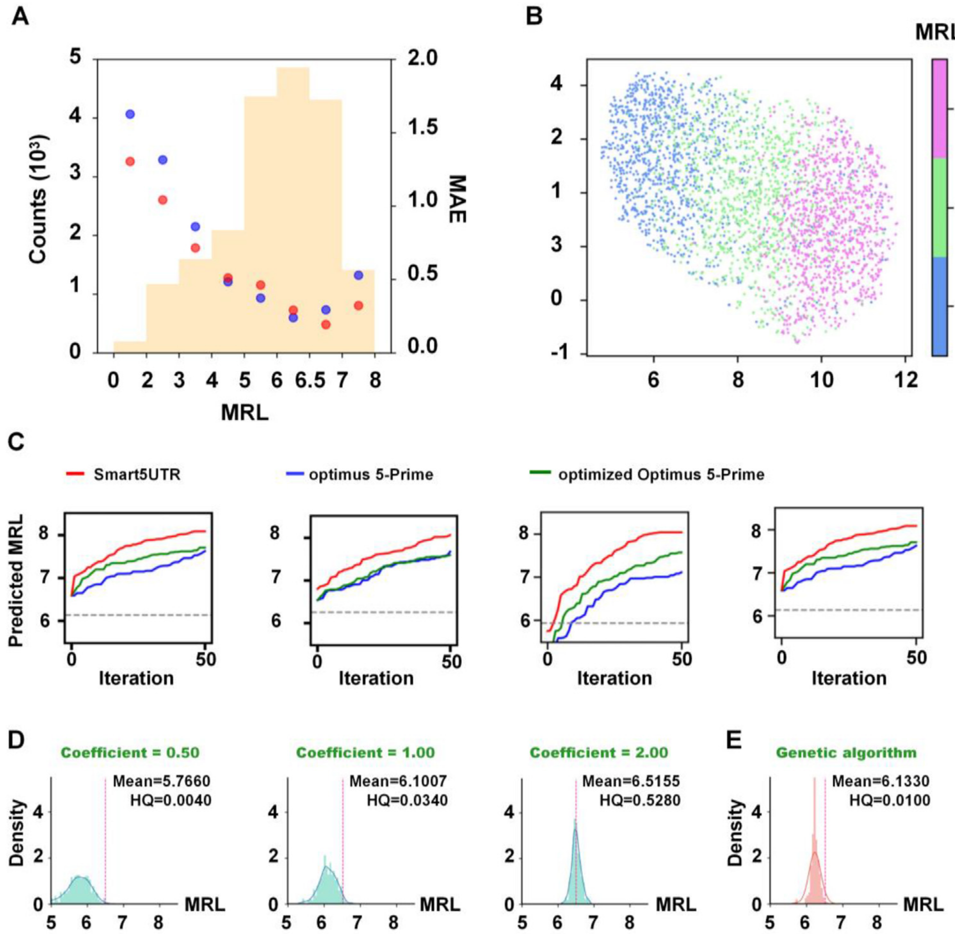


Figure 4 Evaluation of Smart5UTR and Optimus 5-Prime. (A) Sequences frequency distribution of the test dataset and MAE on each MRL segment for Smart5UTR (red) and Optimus 5-prime (blue). On data segments with MRL values higher than 6.5, Smart5UTR has a minor prediction error compared to Optimus 5-prime. (B) UMAP dimensionality reduction for visualizing the latent vectors of Smart5UTR. From the held-out testing dataset, we randomly sampled 1000 5′ UTR from each of three data intervals, high (7 ± 0.5), medium (5 ± 0.5), and low (3 ± 0.5) MRL, and showed their latent vectors. (C) The MRL of sequences after 50 iterations in Smart5UTR (red line), Optimus 5-Prime (blue line) and optimized Optimus 5-Prime (green line). (D) The impact of the strengthening coefficient on Smart5UTR generation. The high-quality (HQ) ratio indicates the proportion of sequences with a higher MRL value than the prototype. (E) The optimization effect of Optimus 5-Prime.

In the generation task, Smart5UTR outperformed the GA used in Optimus 5-Prime. We randomly sampled 5′ UTR sequences with MRL around 6 and fed them into models for 50 iterations. Smart5UTR rendered higher MRL to the sequences in most cases, even though we accelerated the optimization progress of the Optimus 5-Prime by increasing the mutant probability from 0.5 to 0.7 (Fig. 4C and Supporting Information Fig. S3). Subsequently, we used Smart5UTR and Optimus 5-Prime to optimize 1000 randomly selected sequences under the same iterative conditions. Smart5UTR enabled 98% of the sequences to achieve higher MRL values, while Optimus 5-Prime improved 2% of them (Supporting Information Table S3 and Fig. S4). Furthermore, we randomly sampled a sequence with MRL around 6.5 from the test set and fed it into two models for a single iteration of optimization, which was repeated 500 times. Compared to the sequences generated by the two models, the ones from Smar5UTR achieved a higher average MRL and were more likely to have higher MRLs than the prototype (Fig. 4D and E, Supporting Information Fig. S5). Smart5UTR generates sequences by adding random noise to the latent vectors of the input sequences and then multiplying the vectors by a strengthening coefficient. The strengthening coefficient provided a directional perturbation, with the high strengthening coefficient promoting 5′ UTR optimization (Fig. 4D).

3.3. *Validating the efficacy of Smart5UTR in vitro*

To validate the performance of Smart5UTR, we synthesized m1Ψ-modified mRNAs that contained Smart5UTR-designed 5′ UTRs. These mRNAs encoded enhanced green fluorescent protein (EGFP) and were transfected into the HEK293T cells and DC2.4 cells. The expression of EGFP-mRNA is represented by their mean fluorescence intensity (MFI), which was measured by flow cytometry. The MFI of EGFP-mRNAs were highly correlated with the prediction of Smart5UTR in HEK293T cells ($R^2 = 0.8311$, Fig. 5A). The correlation was slightly weak in DC2.4 cells ($R^2 = 0.7175$, Fig. 5B). It was possibly because the dataset used for Smart5UTR building was produced in HEK293T cells. Nevertheless, the m1Ψ-modified mRNAs with different 5′ UTRs generated by Smart5UTR showed relative translation efficiency and Smart5UTR is credible *in vitro*.

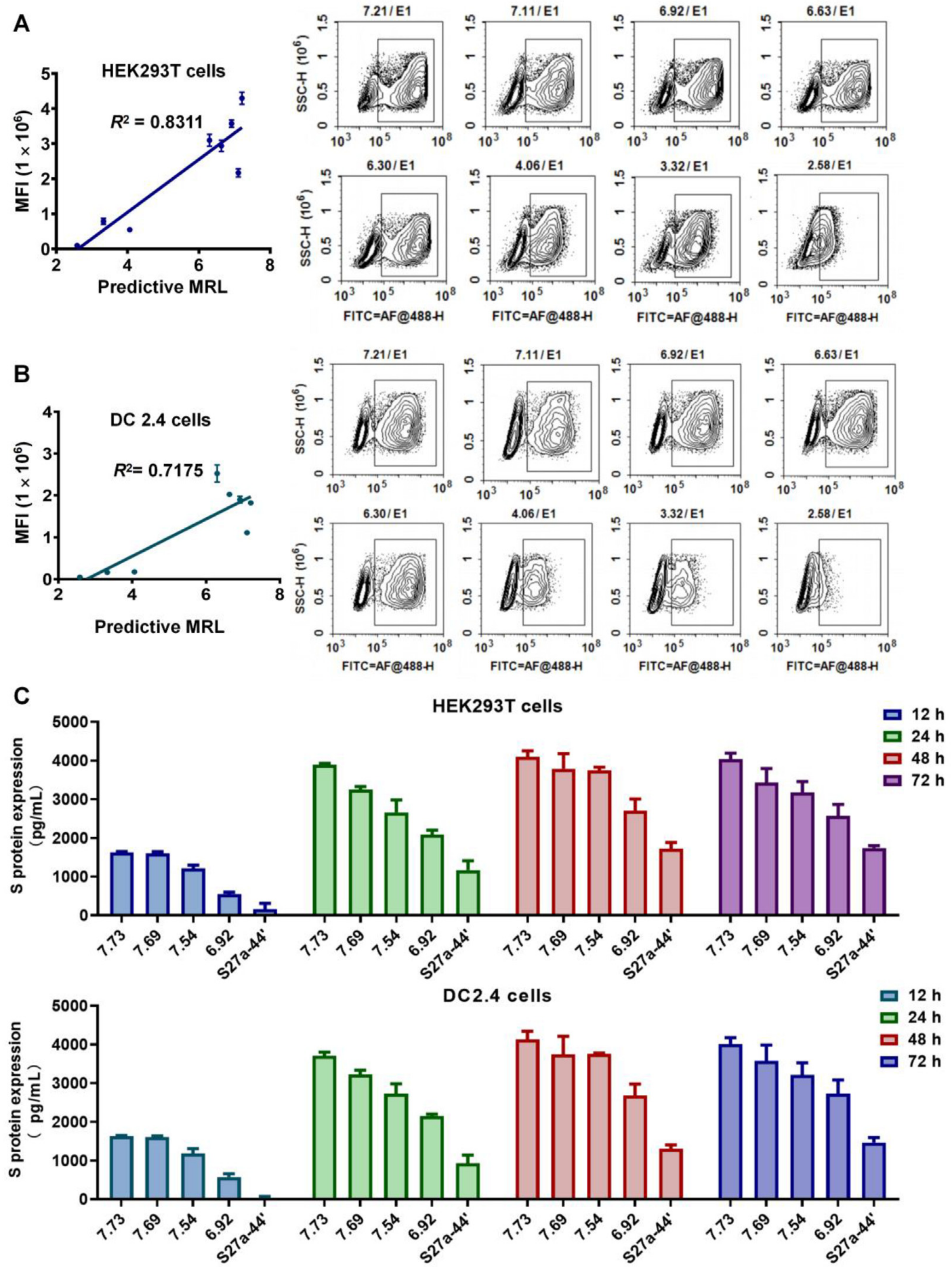


Figure 5 Validation of Smart5UTR *in vitro*. (A) The eGFP mRNAs with different Smart5UTR-predicted MRLs were transfected into the HEK293T cells. (B) The eGFP mRNAs with different Smart5UTR-predicted MRLs transfected into DC2.4 cells. (C) Expression of the S protein of mRNA vaccines in HEK293T cells and DC2.4 cells. The protein level of the spike was assessed by quantitative ELISA after 12, 24, 48 and 72 h of transfection of the COVID-19 mRNA vaccines. All data are presented as mean ± SD ($n = 3$), non-significant (ns), $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ and $^{****}P < 0.0001$. EGFP, enhanced green fluorescent protein; MFI, mean fluorescence intensity; S protein, spike protein.

Furthermore, we used Smart5UTR to *de novo* design superior 5′ UTRs for COVID-19 vaccine, with predicted MRLs of 7.73, 7.69, 7.54 and 6.92 (7.73 5′ UTR, 7.69 5′ UTR, 7.54 5′ UTR and 6.92 5′ UTR). The S27a-44′ 5′ UTR, an endogenous superior 5′ UTR, was used as a positive control, with a predictive MRL value of 5.45[9]. To engineer the mRNA vaccine against SARS-CoV-2, we chose spike glycoprotein as the target antigen based on its outstanding antigenicity and critical function in mediating SARS-CoV-2 entry into host cells[29]. Consequently, we transcribed m1Ψ-modified mRNAs that contained 5′ UTRs mentioned above and encoded Spike protein from the SARS-CoV-2 Delta variant (B.1.617.2). The mRNAs were transfected to HEK293T cells and DC2.4 cells, and the protein was measured by quantitative ELISA after 12, 24, 48 and 72 h of transfection of the COVID-19 mRNA vaccines. We observed that the expression of all mRNAs peaked at 48 h and remained at 72 h. And Smart5UTR-designed superior 5′ UTR rendered higher S protein expression than the endogenous superior 5′ UTR, S27a-44′ UTR (Fig. 5C). In addition, mRNA stability of all COVID-19 vaccines was assessed by fluorogenic quantitative PCR. The stability of the mRNAs was irrelevant to the expression level (Supporting Information Fig. S6). These results indicated that Smart5UTR-designed m1Ψ−5′ UTR enhanced the translation efficacy of mRNAs but not mRNA stability, leading to an improvement in the expression level of COVID-19 vaccines.

3.4. *Smart5UTR-designed superior m1Ψ−5′ UTR benefited SARS-CoV-2 vaccines in mice*

To further verify Smart5UTR *in vivo*, we prepared COVID-19 vaccines with mRNAs that encoded spiked protein and contained 7.73 5′ UTR, 7.69 5′ UTR, 7.54 5′ UTR or 6.92 5′ UTR. These mRNAs were encapsulated in lipid nanoparticles (LNPs), respectively, and were assessed in size and zeta potential (Fig. 6A and B). Mice were administered 5 μg of mRNA by intramuscular injection on Days 0 and 14, and serum was collected on Day 28 for evaluation of antibody response by ELISA (Fig. 6C). The mRNA vaccines with Smart5UTR-designed 5′ UTR evoked stronger humoral immunity than the one containing the S27a-44′ 5′ UTR. The binding antibody titer of the mRNA vaccine containing 7.73, 7.69, 7.54 and 6.92 5′ UTR was 327,680, 204,800, 133,120, and 74,240, respectively, when using the RBD of the SARS-CoV-2 Delta variant (B.1.617.2) as the coating antigen. The titer of the one with S27a-44′ 5′ UTR was low, with merely 2720 (Fig. 6D). All vaccines had impaired titer against the SARS-CoV-2 Omicron variant (B.1.1.529). Nevertheless, the titer of the vaccine with 7.73 5′ UTR was 102,400, and the titer of the vaccine with S27a-44′ 5′ UTR was 2560 (Fig. 6E). The Smart5UTR-designed 5′ UTRs enabled the COVID-19 mRNA vaccine to elicit significantly stronger antibody responses against SARS-CoV-2 Delta and Omicron variants, compared with the one using endogenous superior 5′ UTR, S27a-44′ 5′ UTR.

In response to the unprecedented scale of COVID-19 mRNA vaccinations, the public and providers remain vigilant about vaccine safety. Myocarditis is a rare complication of COVID-19 mRNA vaccination[30], and liver injury following COVID-19 mRNA vaccination has been reported[31,32]. The Smart5UTR-optimized mRNA vaccine showed outstanding safety in a series of evaluations. To

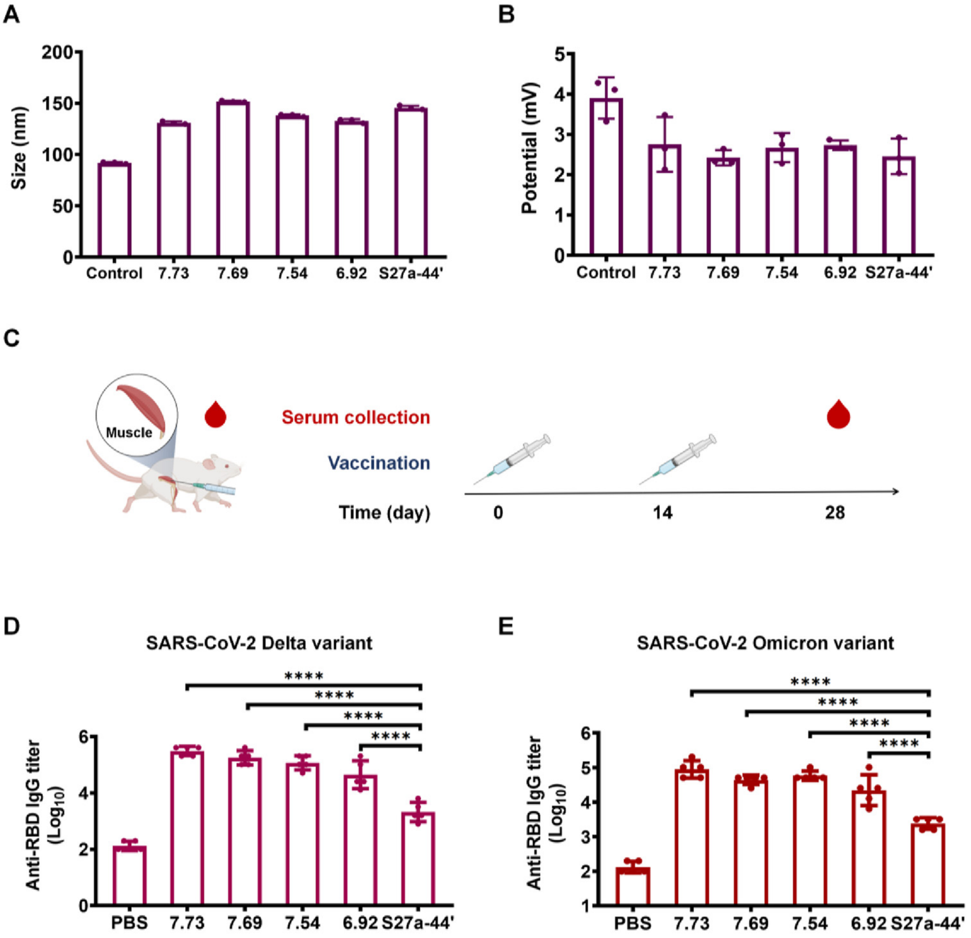


Figure 6 Smart5UTR improved COVID-19 mRNA vaccines in mice. The size (A) and zeta potential (B) of COVID-19 mRNA vaccines ($n = 3$) and the control LNP without loading mRNA. (C) The vaccination schedule. BALB/c mice ($n = 5$) were intramuscularly injected with 5 μg of COVID-19 mRNA vaccines on Days 0 and 14. COVID-19 mRNA vaccines contained 5′ UTRs designed by Smart5UTR or S27a-44′ 5′ UTR. Control mice were injected intramuscularly with PBS. The serum was collected on Day 28 for antibody assay. (D) The endpoint of the IgG titer against RBD of the SARS-CoV-2 Delta variant. (E) The endpoint of the IgG titer against RBD of the SARS-CoV-2 Omicron variant. All data are presented as mean ± SD, non-significant (ns), $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ and $^{****}P < 0.0001$.

assess cardiac safety, we examined the serum creatine kinase MB (CKMB) isoenzyme level and observed no significant elevation. The heart section also indicated the good safety of Smart5UTR-optimized mRNA vaccines (Fig. 7A and Supporting Information Fig. S7). For evaluating hepatic safety, we tested alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), total protein (TP) and albumin (ALB), and all biochemical indices were normal (Fig. 7A). All liver sections did not show vaccination-associated liver injury (Supporting Information Fig. S7). We also tested the serum creatinine (CRE) and urea (URE), which did not show renal injury. No toxicity of Smart5UTR-designed mRNA vaccines was observed in the HE-stained sections of the kidney, lung, and spleen. Moreover, the body weight changes showed insignificant difference between the mice administrated with mRNA vaccines or PBS (Fig. 7B). In conclusion, Smart5UTR-designed mRNA vaccines have good safety.

4. Discussion

The mRNA technology has accelerated the development of COVID-19 vaccines that have effectively and timely protected vulnerable populations from severe and fatal COVID-19. Notably, during recent decades, pandemics have occurred more frequently, and there is an urgent need to improve the vaccine platform to better protect against the new pandemics[1,2]. Therefore, it is necessary to develop potent vaccine platforms and prepare for future pandemics. The mRNA technique has outstanding potential for combating the emerging pandemic due to its rapid development, easy manufacture and convenient scale-up, and thus, deserves further optimization[33,34]. However, no specialized tool exists for directing the sequence design of m1Ψ−mRNA, which plays a key role in mRNA vaccines. To address this issue, we developed a deep generative model, Smart5UTR, to specifically design m1Ψ−mRNA vaccines.

In this research, we proposed that nucleotide modification could alter the superior sequence of mRNA, and observed the difference in sequence preference between m1Ψ-modified and unmodified mRNA based on the analysis of an open dataset. This finding highlighted the need for a dedicated tool for m1Ψ−mRNA design. To develop a practical tool for m1Ψ−mRNA design, we built a deep generation model, Smart5UTR, to search for superior 5′ UTR sequences among numerous potential base combinations. We customized a loss function and network structure, which made Smart5UTR overcome the limitation of the existing model and achieve directional optimization. Smart5UTR has high prediction performance and reconstruction accuracy, particularly for high translation efficacy mRNAs that are vital for vaccine design. The function of Smart5UTR was validated *in vitro*, that Smart5UTR can design 5′ UTRs with target translation efficiency and superior 5′ UTR sequences. To further evaluate Smart5UTR, we used the Smart5UTR-designed 5′ UTR to engineer m1Ψ−mRNA for the COVID-19 vaccine, and these vaccines showed high efficacy in mice. The Smart5UTR-designed superior 5′ UTR helped COVID-19 vaccine to induce strong immune response against the

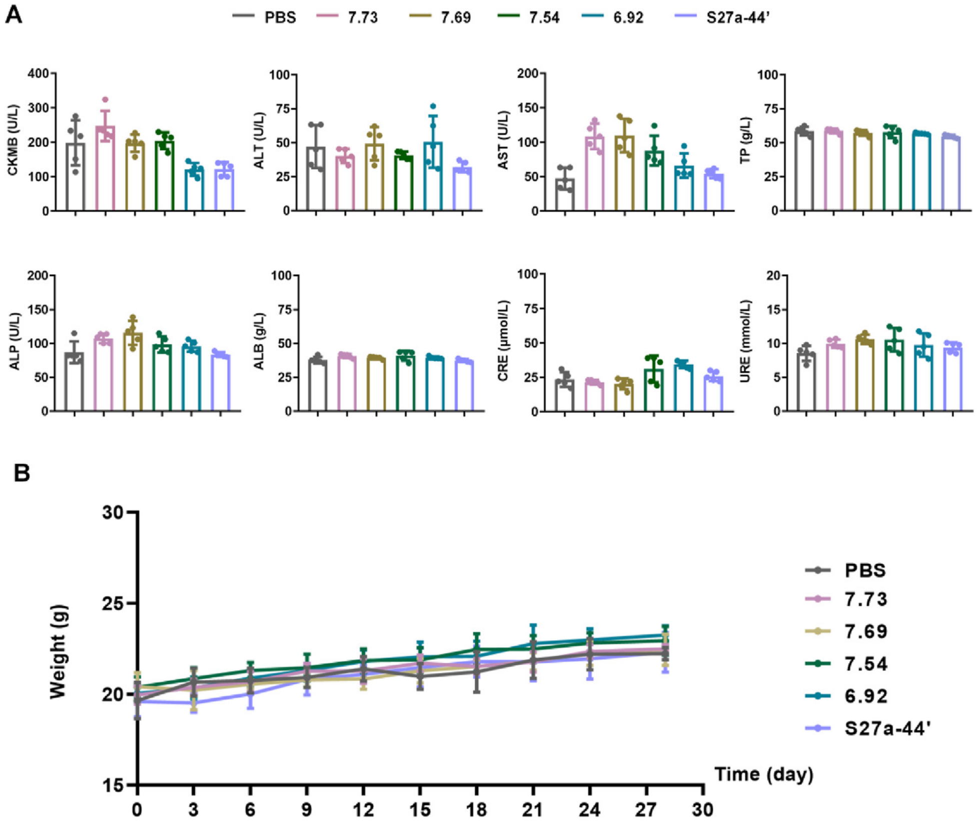


Figure 7 The safety evaluation of COVID-19 mRNA vaccines with Smart5UTR-designed 5′ UTR. (A) Results of the blood biochemical analysis. After two months of booster vaccination, serum from BALB/c mice was collected to analyze CKMB, ALT, AST, ALP, TP, ALB, CRE and URE. (B) The body weight changes of the mice with vaccination or PBS. All data are presented as mean ± SD (n = 5). CKMB, creatine kinase isoenzyme MB; ALT, alanine aminotransferase; AST, aspartate aminotransferase; ALP, alkaline phosphatase; TP, total protein; ALB, albumin; CRE, serum creatinine and urea.

Delta and Omicron variants of SARS-CoV-2 compared to the one with superior endogenous 5′ UTR[9].

We provided a useful tool for m1*Ψ*−5′ UTR design and validated it *in vitro* and *in vivo*, while several issues need to be addressed in further detail. Smart5UTR exhibited satisfactory performance in the high-MRL segment in HEK293T cells, yet instability in DC2.4 cells. This variance was possibly caused by the model's training data, which was produced in HEK293T cells, while translation efficiency of the same mRNA has a certain degree variation in different cell types[35,36]. The applicability range of Smart5UTR is 2−8, where the prediction error of Smart5UTR is accessible (Fig. 4A), while the fluctuation of translation in different cell types should be notified. Moreover, the dataset used for Smart5UTR has an unbalanced distribution of sequences across different MRL value intervals, leading to suboptimal prediction performance in under-sampled high MRL value regions. This limitation also affects the efficiency of the generative algorithm in designing novel 5′ UTR sequences. However, the training dataset used in the research represents the highest quality one available for translation efficiency of m1*Ψ*−mRNA. Despite the challenges, the MTAE structure can boost performance by applying data augmentation to generate UTR sequences. Smart5UTR designs novel high-expression 5′ UTRs, allowing dataset reconstruction. We recommend adding high-MRL 5′ UTRs from Smart5UTR, fixing flawed sequences, and down-sampling low-MRL 5′ UTRs. Training a new MTAE with this resampled dataset may improve prediction and generation of high-MRL 5′ UTRs. We propose an iterative process of data augmentation and model evaluation to retain the best-performing MTAE. Validating machine learning approaches for mRNA sequence design both *in vitro* and *in vivo* represents another challenge. One issue lies in the fluctuating correlation between Smart5UTR-designed 5′ UTRs and their target translational efficiency, which may impact the deep learning model's reliability. Although evaluating the model using a large number of Smart5UTR-designed 5′ UTRs for mRNA drug development and experimental verification can enhance its reliability, this approach is costly and time-consuming. Furthermore, the performance of deep learning models can vary across *in silico*, *in vitro*, and *in vivo* environments, complicating their applications in the design of clinically effective mRNA drugs. Considering these limitations, future research should focus on refining the model along with the dataset and addressing these challenges to improve the potential of machine learning-guided mRNA vaccine design.

5. Conclusions

This study pioneered the *de novo* design of m1*Ψ*−5′ UTR based on machine learning. Our newly developed deep-generative model, Smart5UTR, displayed several advantages over the existing models and the conventional methods, as expected. In comparison to previous models, Smart5UTR more accurately predicts and effectively optimizes 5′ UTR sequences with high translation efficiency. Thus, it can be used to design superior 5′ UTRs of m1*Ψ*-modified mRNA for mRNA vaccines or other mRNA therapeutics, especially for those requiring high protein expression, such as protein replacement treatment.

Acknowledgments

This work was financially supported by the Postdoctoral Research Foundation of National Key S&T Special Projects (2018ZX09201018-024, China), Sichuan Province Science and Technology Support Program (2022YFH0001 and 2021YFH0003, China). The author would like to thank Prof. Zhen Gu (Zhejiang University, Zhejiang, China) for kindly providing DC2.4 cells. The author would like to thank Xiaojiao Wang (Core Facilities of West China Hospital for technical assistance, Chengdu, China) and Yi Zhang (Core Facilities of West China Hospital for technical assistance, Chengdu, China). Some elements in graphical abstract, Figs. 1 and 6 are from BioRender.com (2022), retrieved from https://app.biorender.com/. We originally designed all the figures.

Author contributions

Xiaoshan Tang and Miaozhe Huo designed the research, carried out the experiments, performed data analysis, wrote and revised the manuscript. Yuting Chen carried out the experiments and revised the manuscript. Shugang Qin, Na Fan, Zhongshan He and Xi He revised the manuscript. Jiaqi Luo, Xin Jiang, Yongmei Liu, Xing Duan, Ruohan Wang, Lingxi Chen and Hao Li participated part of the experiments. Xiangrong Song, Shuai Cheng Li, Bairong Shen and Hai Huang designed the research, acquired the research funding, and revised the manuscript. All of the authors have read and approved the final manuscript.

Conflicts of interest

The authors declare no conflict of interest.

Appendix A. Supporting information

Supporting data to this article can be found online at https://doi.org/10.1016/j.apsb.2023.11.003.

References

1. Excler JL, Saville M, Berkley S, Kim JH. Vaccine development for emerging infectious diseases. *Nat Med* 2021;**27**:591−600.
2. Graham BS, Sullivan NJ. Emerging viral diseases from a vaccinology perspective: preparing for the next pandemic. *Nat Immunol* 2018;**19**: 20−8.
3. Chaudhary N, Weissman D, Whitehead KA. mRNA vaccines for infectious diseases: principles, delivery and clinical translation. *Nat Rev Drug Discov* 2021;**20**:817−38.
4. Pardi N, Hogan MJ, Porter FW, Weissman D. mRNA vaccines—a new era in vaccinology. *Nat Rev Drug Discov* 2018;**17**:261−79.
5. Barbier AJ, Jiang AY, Zhang P, Wooster R, Anderson DG. The clinical progress of mRNA vaccines and immunotherapies. *Nat Biotechnol* 2022;**40**:840−54.
6. Hinnebusch AG, Ivanov IP, Sonenberg N. Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science* 2016;**352**: 1413−6.
7. Qin S, Tang X, Chen Y, Chen K, Fan N, Xiao W, et al. mRNA-based therapeutics: powerful and versatile tools to combat diseases. *Signal Transduct Targeted Ther* 2022;**7**:166.
8. 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* 2018;**19**:158−74.
9. Zeng C, Hou X, Yan J, Zhang C, Li W, Zhao W, et al. Leveraging mRNA sequences and nanoparticles to deliver SARS-CoV-2 antigens *in vivo*. *Adv Mater* 2020;**32**:e2004452.
10. Ferizi M, Aneja MK, Balmayor ER, Badieyan ZS, Mykhaylyk O, Rudolph C, et al. Human cellular CYBA UTR sequences increase mRNA translation without affecting the half-life of recombinant RNA transcripts. *Sci Rep* 2016;**6**:39149.

11. Trepotec Z, Aneja MK, Geiger J, Hasenpusch G, Plank C, Rudolph C. Maximizing the translational yield of mRNA therapeutics by minimizing 5′-UTRs. *Tissue Eng* 2019;**25**:69−79.
12. Svitkin YV, Cheng YM, Chakraborty T, Presnyak V, John M, Sonenberg N. *N*1-Methyl-pseudouridine in mRNA enhances translation through eIF2alpha-dependent and independent mechanisms by increasing ribosome density. *Nucleic Acids Res* 2017;**45**:6023−36.
13. Svitkin YV, Gingras AC, Sonenberg N. Membrane-dependent relief of translation elongation arrest on pseudouridine- and *N*1-methyl-pseudouridine-modified mRNAs. *Nucleic Acids Res* 2022;**50**:7202−15.
14. Dolgin E. Trial settles debate over best design for mRNA in COVID vaccines. *Nature* 2023;**613**:419−20.
15. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. *Nat Rev Mol Cell Biol* 2022;**23**:40−55.
16. Amin N, McGrath A, Chen YPP. Evaluation of deep learning in non-coding RNA classification. *Nat Mach Intell* 2019;**1**:246−56.
17. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* 2018;**15**:20170387.
18. Sample PJ, Wang B, Reid DW, Presnyak V, McFadyen IJ, Morris DR, et al. Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nat Biotechnol* 2019;**37**:803−9.
19. Cao J, Novoa EM, Zhang Z, Chen WCW, Liu D, Choi GCG, et al. High-throughput 5′ UTR engineering for enhanced protein production in non-viral gene therapies. *Nat Commun* 2021;**12**:4138.
20. Fan N, Chen K, Zhu R, Zhang Z, Huang H, Qin S, et al. Manganese-coordinated mRNA vaccines with enhanced mRNA expression and immunogenicity induce robust immune responses against SARS-CoV-2 variants. *Sci Adv* 2022;**8**:eabq3500.
21. Chen K, Fan N, Huang H, Jiang X, Qin S, Xiao W, et al. mRNA vaccines against SARS-CoV-2 variants delivered by lipid nanoparticles based on novel ionizable lipids. *Adv Funct Mater* 2022;**32**:2204692.
22. Shi H, Chai P, Jia R, Fan X. Novel insight into the regulatory roles of diverse RNA modifications: re-defining the bridge between transcription and translation. *Mol Cancer* 2020;**19**:78.
23. Cuperus JT, Groves B, Kuchina A, Rosenberg AB, Jojic N, Fields S, et al. Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. *Genome Res* 2017;**27**:2015−24.
24. Zhang H, Wang Y, Wu X, Tang X, Wu C, Lu J. Determinants of genome-wide distribution and evolution of uORFs in eukaryotes. *Nat Commun* 2021;**12**:1076.
25. Hernandez G, Osnaya VG, Perez-Martinez X. Conservation and variability of the AUG initiation codon context in eukaryotes. *Trends Biochem Sci* 2019;**44**:1009−21.
26. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res* 2004;**14**:1188−90.
27. Aoki G, Sakakibara Y. Convolutional neural networks for classification of alignments of non-coding RNA sequences. *Bioinformatics* 2018;**34**:i237−44.
28. Gruber AR, Lorenz R, Bernhart SH, Neuboock R, Hofacker IL. The Vienna RNA websuite. *Nucleic Acids Res* 2008;**36**:W70−4.
29. Bok K, Sitar S, Graham BS, Mascola JR. Accelerated COVID-19 vaccine development: milestones, lessons, and prospects. *Immunity* 2021;**54**:1636−51.
30. Bozkurt B, Kamat I, Hotez PJ. Myocarditis with COVID-19 mRNA vaccines. *Circulation* 2021;**144**:471−84.
31. Shroff H, Satapathy SK, Crawford JM, Todd NJ, VanWagner LB. Liver injury following SARS-CoV-2 vaccination: a multicenter case series. *J Hepatol* 2022;**76**:211−4.
32. Efe C, Kulkarni AV, Terziroli Beretta-Piccoli B, Magro B, Stättermayer A, Cengiz M, et al. Liver injury after SARS-CoV-2 vaccination: features of immune-mediated hepatitis, role of corticosteroid therapy and outcome. *Hepatology* 2022;**76**:1576−86.
33. Corbett KS, Edwards DK, Leist SR, Abiona OM, Boyoglu-Barnum S, Gillespie RA, et al. SARS-CoV-2 mRNA vaccine design enabled by prototype pathogen preparedness. *Nature* 2020;**586**:567−71.
34. Vogel AB, Kanevsky I, Che Y, Swanson KA, Muik A, Vormehr M, et al. BNT162b vaccines protect rhesus macaques from SARS-CoV-2. *Nature* 2021;**592**:283−9.
35. Genuth NR, Barna M. Heterogeneity and specialized functions of translation machinery: from genes to organisms. *Nat Rev Genet* 2018;**19**:431−52.
36. Buszczak M, Signer RA, Morrison SJ. Cellular differences in protein synthesis regulate tissue homeostasis. *Cell* 2014;**159**:242−51.