mRNA vaccine sequence and structure design and optimization: Advances and challenges

Received for publication, September 26, 2024, and in revised form, November 13, 2024 Published, Papers in Press, November 26, 2024,
https://doi.org/10.1016/j.jbc.2024.108015

Lei Jin[1,‡], **Yuanzhe Zhou**[1,‡], **Sicheng Zhang**[1,‡], and **Shi-Jie Chen**[1,2,*]

From the [1]*Department of Physics and Astronomy, University of Missouri, Columbia, Missouri, USA;* [2]*Department of Biochemistry, MU Institute for Data Science and Informatics, University of Missouri, Columbia, Missouri, USA*

Reviewed by members of the JBC Editorial Board. Edited by Karin Musier-Forsyth

Messenger RNA (mRNA) vaccines have emerged as a powerful tool against communicable diseases and cancers, as demonstrated by their huge success during the coronavirus disease 2019 (COVID-19) pandemic. Despite the outstanding achievements, mRNA vaccines still face challenges such as stringent storage requirements, insufficient antigen expression, and unexpected immune responses. Since the intrinsic properties of mRNA molecules significantly impact vaccine performance, optimizing mRNA design is crucial in preclinical development. In this review, we outline four key principles for optimal mRNA sequence design: enhancing ribosome loading and translation efficiency through untranslated region (UTR) optimization, improving translation efficiency *via* codon optimization, increasing structural stability by refining global RNA sequence and extending in-cell lifetime and expression fidelity by adjusting local RNA structures. We also explore recent advancements in computational models for designing and optimizing mRNA vaccine sequences following these principles. By integrating current mRNA knowledge, addressing challenges, and examining advanced computational methods, this review aims to promote the application of computational approaches in mRNA vaccine development and inspire novel solutions to existing obstacles.

Vaccination has been the most significant health intervention in reducing mortality since 1974 (1). The World Health Organization estimates that global vaccination has prevented 154 million deaths over the past 50 years (1). Vaccines are designed to train the immune system to recognize antigens produced by pathogens or malignant cells (2–5) and stimulate a robust adaptive immune response against infections or cancers (6, 7). Recent advances in messenger RNA (mRNA) technology have enabled the delivery of antigen-encoding mRNA molecules into host cells, facilitating antigen expression within the human body (4). Compared to conventional vaccines (*e.g.*, inactivated, live-attenuated, and subunit/recombinant/polysaccharide/conjugate types), mRNA vaccines offer several distinct advantages. First, mRNA can be designed and manufactured in a rapid, scalable, and cost-effective manner owing to the high yields of cell-free *in vitro* transcription (IVT) technique (4, 8). For example, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) mRNA vaccine mRNA-1273 entered clinical testing just 2 months after the genetic sequence of SARS-CoV-2 was identified (9, 10). Second, the mRNA vaccine is free of infectious pathogens, eliminating the risk of infection upon vaccination (4). Additionally, the cell-free, *in vitro* manufacturing environment minimizes safety concerns typically associated with other platforms, such as cell-derived impurities and viral contaminants (11). Third, the mRNA vaccine can encode multiple specific antigens in a single formulation, allowing for both precise and robust immune responses against resilient pathogens (8, 12).

Despite the remarkable success of mRNA vaccines in coronavirus disease 2019 (COVID-19) prevention, several challenges remain for their broader applications, particularly in cancer treatment—the field that initially sparked their development (13–15). One significant hurdle is the insufficient antigen-expressing efficiency of artificial mRNAs used in vaccines, which often fails to elicit a long-lasting immune response, necessitating multiple booster injections to establish effective protection (16, 17). Another obstacle is the intrinsic thermal instability of mRNA molecules, requiring low-temperature storage and posing logistical challenges that can impede vaccine distribution during a pandemic (17). Furthermore, the reactogenicity[1] of artificial mRNA and its delivery carriers, along with potential undesirable protein translation, may lead to unexpected side effects or even hypersensitivity reactions (18–20).

Advances in mRNA molecular and metabolic biology have led to potential solutions for some of the challenges in designing effective mRNA vaccines (21, 22). These approaches include optimizing mRNA sequences and structures to enhance ribosome loading and translation efficiency (TE), increasing mRNA thermal stability by designing sequences that fold into more stable structures, reducing unstructured

‡ These authors contributed equally to this work.
* For correspondence: Shi-Jie Chen, chenshi@missouri.edu.

[1] Reactogenicity refers to the ability of a vaccine to elicit expected short-term adverse reactions or inflammatory responses following immunization. Reactogenic events can include local effects, such as pain, redness, swelling, or induration at the injection site, as well as systemic effects like myalgia, fever, and headache. Ideally, vaccines should induce a robust immune response while minimizing reactogenicity.

RNA local motifs that may trigger cellular immune responses or attract in-cell ribonucleases, as well as avoiding slippery sequence motifs and downstream frameshift structural signals to maintain translation fidelity. These strategies collectively aim to improve the efficacy and stability of mRNA-based vaccines. Since the properties of an mRNA vaccine are largely determined by its primary sequence, most of the current efforts to improve efficacy focus on sequence design and optimization (21).

In eukaryotic cells, precursor mRNAs (pre-mRNAs) produced by transcription need to undergo processing to form mature mRNAs before they are ready for translation during protein synthesis. This maturation process typically involves three main steps: 5′ end capping, 3′ end polyadenylation, and intron splicing. In order to have an improved expression efficiency and translation fidelity, artificially designed mRNAs used for vaccines should ideally be mature and ready for translation before entering host cells (14, 15, 23). Since the IVT technique can synthesize RNA molecules from DNA templates in a controlled laboratory environment, it simplifies production and maturation of custom-designed mRNA sequences. This method retains only the regulatory elements essential for protein synthesis, allowing researchers to design, evaluate, and optimize the functions of each individual mRNA segment separately to achieve higher expression efficiency, in-solution/cell stability, and expression fidelity (21, 22, 24). The simplified mRNA produced by IVT consists of five components: 5′ cap, 5′ untranslated region (UTR), open reading frame (ORF), 3′ UTR, and 3′ end (*i.e.*, poly(A) tail). The IVT produced mRNA sequence starts with the 5′ cap, usually modified guanosine (7-methylguanosine, abbreviated m^7G), whose purpose is to stabilize IVT mRNA against 5′ exonucleolytic degradation and initiate translation within the cell (25–27). Following the 5′ cap, the 5′ UTR contains crucial regulatory regions such as the ribosome binding site (RBS) and internal ribosome entry site (IRES). These elements control translation by recruiting the ribosome and other translation factors (28–30). The ORF, located between the start and stop codons, contains the essential coding sequence (CDS) that encodes the target protein antigen (31–33). Without the need of intron splicing in IVT-produced mRNA, the ORF can simply consist of a single continuous CDS (34). The 3′ UTR, together with the 3′ end, protects the mRNA from enzymatic degradation in the cytoplasm by recruiting poly(A)-binding proteins (PABPs) (29, 31, 35, 36). Figure 1 illustrates a simplified manufacturing procedure and mechanism of action for mRNA vaccines.

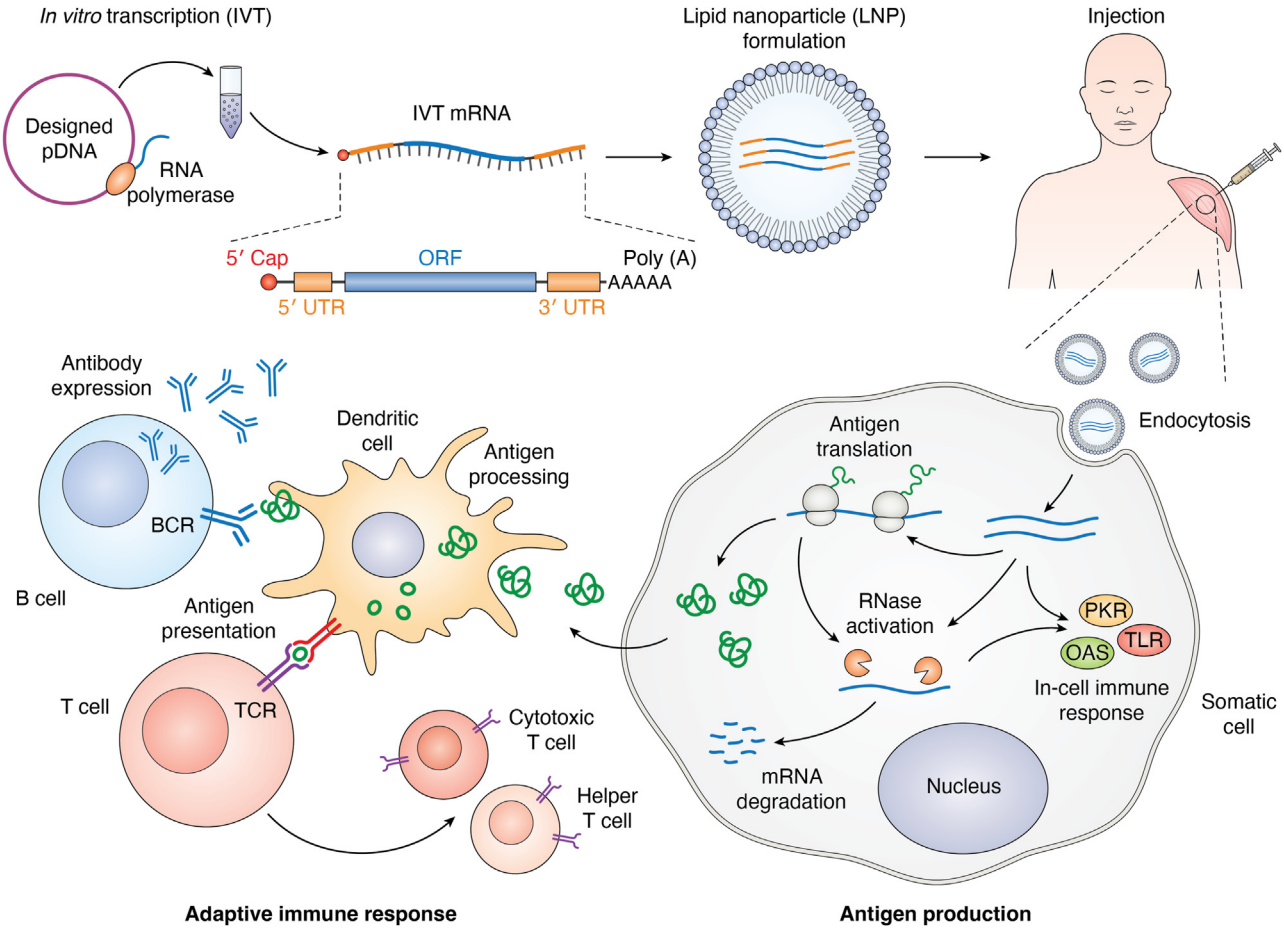

Figure 1. Simplified mRNA vaccine manufacturing procedure and mechanism of action. The artificial mRNAs, designed based on the antigen protein sequence, are *in vitro* transcribed and encapsulated in lipid nanoparticles (LNPs). Upon intramuscular injection, LNPs facilitate mRNAs' entry into muscle cells *via* endocytosis. The translated antigens stimulate adaptive immune responses, activating B and T cells against potential infections. Injected mRNAs are subsequently degraded by cellular ribonucleases (RNases).

Most current computational models for designing and optimizing mRNA sequences focus on individual components of *in vitro* transcribed mRNAs (21). Over the past few years, the 5′ UTR—a major regulatory region of mRNA that determines ribosome loading efficiency—has been investigated (37, 38). Based on these findings, computational models have been proposed to design artificial mRNAs by screening/selecting naturally occurring 5′ UTRs. Additionally, some of the models are also capable of further optimizing/designing 5′ UTR sequences. Meanwhile, several studies have investigated the ORF region containing the CDS (31–33). However, the high variability in codons makes exhausting all possible CDS configurations computationally infeasible. For instance, an ORF encoding a target antigen like the SARS-CoV-2 spike protein could theoretically have over 10^{632} possible CDS variations due to codon diversity (23). While current clinically available mRNA vaccines generally select the optimal codon for each amino acid (23), there remains considerable room for improvement in ORF design when considering other ORF-related factors that impact the vaccine's overall performance. Even if one can identify the optimal sequence for each individual mRNA component, these separately designed segments require validation when integrated into a full mRNA sequence. This step is crucial because long-range interactions between segments and global refolding of the mRNA may lead to unexpected outcomes (39–41).

In this review, we introduce currently available computational approaches for mRNA sequence design and optimization together with their applications in mRNA vaccine development. By addressing challenges in mRNA vaccine development, and exploring advanced computational approaches, we aim to promote the development of computational methods in mRNA vaccine design and inspire new strategies to overcome existing challenges.

Computational mRNA vaccine sequence and structure design and optimization

To maximize the potential of mRNA vaccines, an optimally designed IVT mRNA sequence should achieve high antigen expression efficiency, exhibit low degradation rates both in solution and within cells, and minimize reactogenicity. Based on the functional segmentation of the mRNA sequence, the design objectives for the UTR, ORF, and the overall sequence can be classified into four key principles (see Fig. 2 for an

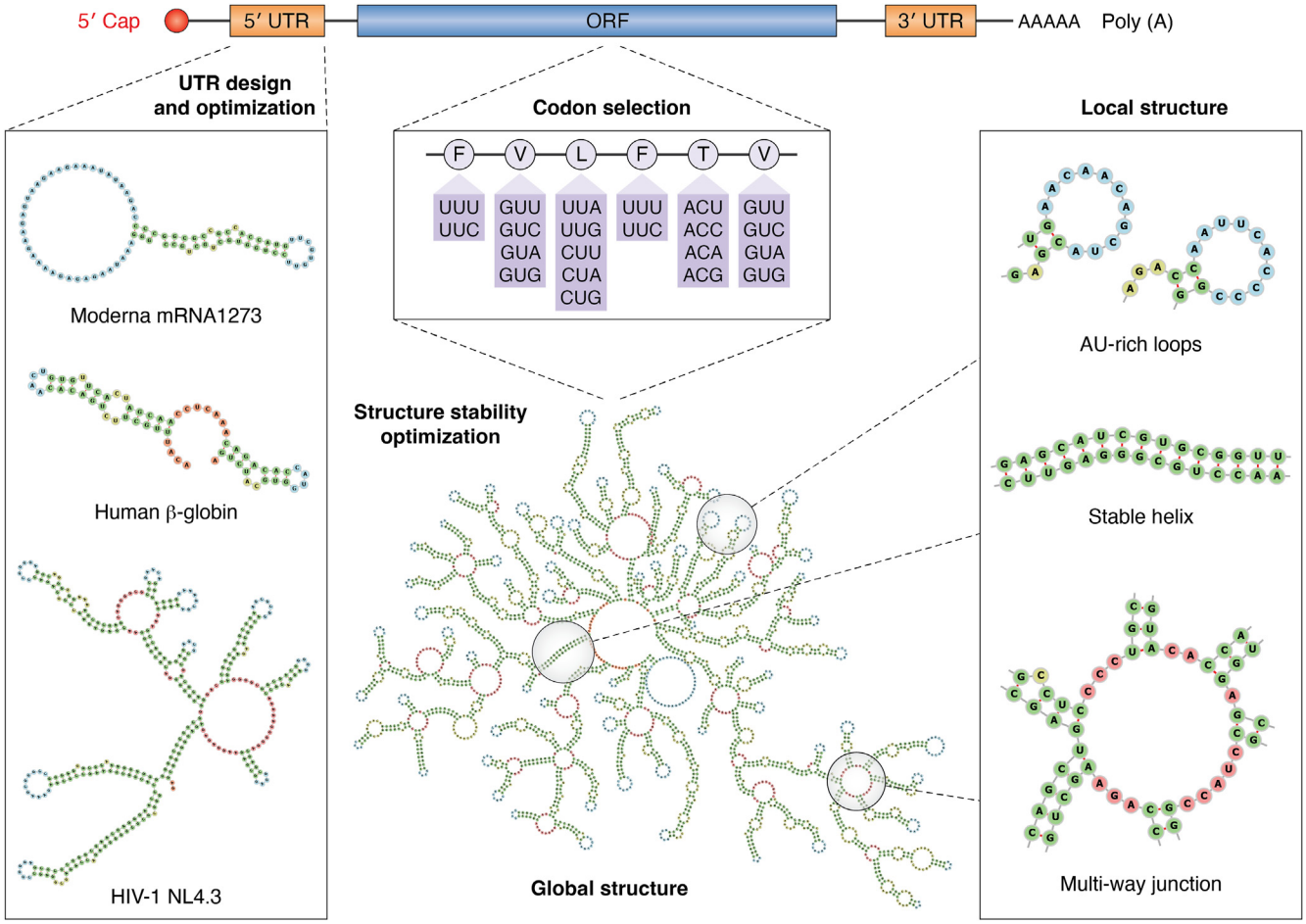

Figure 2. Illustration of *in vitro* transcription (IVT) mRNA components and their corresponding design/optimization processes: 1. Designing and optimizing untranslated region (UTR) sequence; 2. Optimizing open reading frame (ORF) sequence with optimal codon usage; 3. Adjusting and fine-tuning global and local structures. The illustrated 2D structures are generated using RNAfold (112), RNA structure (72), Vfold2D (114), and a landscape-zooming cotranscriptional folding model (183).

illustration of the IVT mRNA components and their corresponding design/optimization processes).

1. Design and optimize UTRs to enhance ribosome loading and translation efficiency.
2. Optimize codon usage in the ORF to improve translation efficiency.
3. Refine the global sequence to adopt a stable structure for increased in-solution stability.
4. Adjust local sequences to avoid specific sequence/structure motifs, thereby extending in-cell lifetime and improving expression fidelity.

To implement these principles effectively, it is essential to be able to evaluate ribosome loading efficiency, predict or measure both global and local 2D/3D structures, and optimize or generate mRNA sequences either in segments or as a whole. Since some principles may involve trade-offs (*e.g.*, optimizing codon usage might conflict with maintaining/increasing structural stability), the final mRNA sequence design must balance these competing objectives to achieve an optimal outcome.

Optimizing and designing UTR sequence to enhance ribosome loading and translation efficiency

5′ UTR

To maximize antigen protein expression in target host cells, artificially designed mRNA should feature a 5′ capped UTR with high ribosome loading efficiency. A straightforward approach for finding optimal 5′ UTR is to select 5′ UTR sequences from existing human mRNAs with known high expression efficiencies in the target cell type. As shown in Figure 2, multiple 5′ UTR candidates can be utilized for the designed mRNA vaccines, including those occurring naturally in various species. However, naturally derived 5′ UTRs may contain various regulatory elements that could also hinder ribosome loading under certain conditions (31, 42–45). For instance, iron response element (IRE) can block ribosomal complex binding under low cellular iron levels (43) and the existence of upstream ORFs (uORFs) can reduce the translation efficiency by regulating the number of ribosomal preinitiation complexes (PICs) that enter ORF region (42). Therefore, only essential regulatory elements that positively promote ribosome loading would be retained when designing 5′ UTR sequences for mRNA vaccine. For instance, the HIV-1 5′ UTR contains over six functional structural motifs that regulate viral mRNA dimerization, splicing, gene expression, viral packaging, and replication. However, the different structural motifs may impact gene expression efficiency in mRNA

Technical terms

Random Forest (RF): An ensemble learning method that constructs multiple decision trees during training and outputs the average prediction (for regression) or majority vote (for classification) of individual trees. Each tree is built using a random subset of the training data and features, which helps prevent overfitting.

Multilayer Perceptron (MLP): A basic type of feedforward artificial neural network that consists of an input layer, output layer, and one or more hidden layers with many neurons stacked together. Neurons in one layer are fully connected to every neuron in the next layer with nonlinear activation functions.

Recurrent Neural Networks (RNN): A class of artificial neural networks designed to work with sequential data by maintaining an internal state (memory). This memory can capture information from previous inputs and update dynamically at each time step based on both current input and previous internal state. The temporal processing capability makes RNNs particularly effective for tasks involving sequences, such as natural language processing, speech recognition, and time series analysis.

Convolutional Neural Networks (CNN): A type of artificial neural network specialized for processing grid-structured data, particularly images. CNNs use convolutional layers with learnable filters to automatically detect and extract hierarchical features from input data. They are particularly effective for computer vision tasks, such as image classification and object detection.

Graph Neural Networks (GNN): A class of artificial neural networks designed to process and analyze graph-structured data, where nodes represent entities and edges capture relationships between them. By learning representations at multiple levels—nodes, edges, and entire graph structures—GNNs excel at tasks ranging from molecular property prediction in chemistry to social network analysis.

Generative Adversarial Networks (GAN): A machine learning framework for generative artificial intelligence where two neural networks compete with each other in the form of a zero-sum game. A generator creates synthetic data, while a discriminator tries to distinguish between real and generated data. Through this competitive process, both networks improve simultaneously: the generator becomes increasingly adept at creating realistic data, while the discriminator becomes better at detecting subtle differences. This architecture has enabled breakthrough applications in image synthesis, art creation, and data augmentation.

Autoencoders: A type of artificial neural network that learns the lower-dimensional representation (encoding) of unlabeled data and then reconstructs it (decoding) from the encoding in an unsupervised fashion. The encoded representation often captures important features of the input data, which makes autoencoders suitable for various applications, such as dimensionality reduction, feature detection, and image generation.

Long Short-Term Memory (LSTM) Networks: A variant of RNNs that aims to mitigate the vanishing gradient problem commonly encountered by traditional RNNs. It uses specialized memory cells with input, output, and forget gates to control information flow. LSTMs are particularly good at learning long-term dependencies in sequential data compared to traditional RNNs.

Gated Recurrent Unit (GRU): A simplified version of LSTM networks that uses update and reset gates to control information flow. GRUs help solve the vanishing gradient problem in standard RNNs while being computationally more efficient than LSTMs.

Bidirectional Encoder Representations From Transformers (BERT): A transformer-based language model that learns to represent text as a sequence of vectors using self-supervised learning. BERT is pre-trained on large text corpora using masked language modeling and next-sentence prediction tasks, then fine-tuned for specific tasks. It is notable for its significant improvement over previous natural language processing models in capturing latent representations and complex contextual relationships in text.

Large Language Model (LLM): A category of foundation models, typically transformer-based, trained on massive amounts of text data to understand and generate human-like text. LLMs can perform various language tasks like translation, summarization, and question-answering without task-specific training. They learn patterns and relationships in language through self-supervised learning on large-scale datasets.

vaccines in distinct ways. To identify and design the most efficient 5′UTRs to enhance mRNA vaccine's gene expression efficiency remains a challenge. 5′ UTRs with essential components—such as the Human β-globin mRNA 5′ UTR, which consists of only two structured motifs—are widely employed in vaccine designs to optimize mRNA vaccines. In the case of mRNA-1273, a novel, designed, and patented 5′ UTR sequence (46), featuring a less complex structure, was utilized to achieve the desired vaccine performance. Further optimization can also be achieved through RNA base mutations or base-pair alterations in the 5′ UTR sequence, which can affect structural stability of the optimized mRNA. Additionally, more advanced approaches involve *de novo* design of the 5′ UTR sequence, where regulatory elements are manually selected and combined to achieve specific goals.

Statistics based on annotations from GENCODE v19 (47) indicate that only 30% of human 5′ UTRs are 100 nucleotides or shorter, with an average length of approximately 200 nucleotides (48). This vast sequence configurational space (*e.g.*, $\sim 4^{200}$) impedes efficient optimization or design of 5′ UTR sequences for mRNA vaccines. In recent years, various machine learning (ML) models have been proposed for modeling the effect of the 5′ UTR on protein expression, as well as its optimization and design. These models range from random forests (RF), convolutional neural networks (CNN), and generative adversarial networks (GAN) to language model (LM)-based approaches (48–57). Generally, these models attempt to predict a variety of experimental quantities based on the given 5′ UTR sequences or their related biological features. Examples of these experimental quantities include mean ribosome load (MRL), a proxy for translation rate that measures the average number of ribosomes associated with a given RNA (49), and translation efficiency (TE), a ratio that quantifies the rate at which mRNA is translated into proteins (50). Studies often combine these models with various optimization algorithms to find specific 5′ UTR sequences that achieve improved translation efficiency. In the following, we briefly review recent ML approaches with a focus on their specific applications rather than technical details of the models.

To investigate the effect of different 5′ UTR sequence variants on translation, Sample *et al.* (49) trained a CNN based model using MRL values obtained from a massively parallel reporter assay (MPRA) based polysome profiling experiment. The experiment used a library of 280,000 synthetic gene sequences, each containing a 50-nt fully random sequence inserted between a defined upstream 25-nt sequence for polymerase chain reaction (PCR) amplification and a downstream enhanced green fluorescent protein (eGFP) coding sequence. The proposed CNN model, termed Optimus 5-Prime, consists of three convolutional layers and a fully connected layer. It takes one-hot[2] encoded 5′ UTR sequences as input and predicts their MRLs. Trained on 260,000 sequences, the model could explain 93% of the MRL variation (measured by the square of the Pearson correlation coefficient) in the test set with 20,000 sequences, a significant improvement over various position-specific k-mer linear models (49). The same experiment was also performed on a synthetic 5′ UTR library derived from ClinVar database (58) that has approximately 25,000 UTRs consisting of the first 50 nucleotides upstream of the start codon of 22,747 common and 2253 variant human 5′ UTR sequences. The model could explain 82% of the observed MRL variation and identify the effect of various variants on ribosome loading. Further evaluation showed the model's ability to predict MRL for modified mRNAs containing pseudouridine/1-methyl pseudouridine, albeit with lower performance. Coupled with a genetic algorithm, the authors also demonstrated the model's ability to design new 5′ UTRs targeting specific levels of protein expression. A major drawback of the model is the lack of flexibility in modeling 5′ UTRs of varying length, as the fixed-length input requires longer sequences to be truncated, resulting in information loss (48).

As the majority (∼70%) of human 5′ UTRs are longer than 100 nt (48), Optimus 5-Prime, trained with a fixed-length input (*e.g.*, 100-nt), cannot generalize well to longer sequences. To alleviate this issue, Karollus *et al.* (48) proposed a similar CNN model, termed FramePool, that extends the capabilities of Optimus 5-Prime to handle 5′ UTRs of any length. The key idea behind FramePool is translation frame-dependent pooling, where convolutional outputs are separated according to the underlying biological reading frame, and global pooling is performed for each frame separately. When trained and tested on the same dataset of 280,000 sequences containing 50-nt random synthetic sequence in the 5′ UTR, both Optimus 5-Prime and FramePool exhibited similar performance in predicting MRL values. However, when applying both trained models to a test set consisting of 7600 sequences with lengths ranging from 25 to 100 nt (49), the Pearson correlation coefficient dropped to 0.743 for Optimus 5-Prime while achieving 0.901 for FramePool. This shows that the frame pooling technique enables the model to treat sequences of variable length and allows for effective generalization to 5′ UTR sequences considerably longer than those used in training. Nevertheless, the FramePool-predicted MRLs showed considerably lower correlation (ranging from 0.11 to 0.25 for Pearson correlation coefficient) with the experimentally measured TEs when tested on six datasets containing human transcripts (59–64). This is possibly due to the fact that both the CDS and the 3′ UTR, which have been shown to affect the ribosome load and protein-to-mRNA ratios (PTR) (63–65), vary between transcripts in endogenous data. It was also found that FramePool-predicted effects on MRL for single nucleotide variants (SNVs) correlated with the level of phylogenetic conservation at that position, a trend especially pronounced for nucleotides within the 100 nt of the canonical start codon. Furthermore, integrating the models with the Kipoi framework (66), an API and repository of ready-to-use trained models for genomics, enables practitioners to efficiently analyze any human 5′ UTR variant or mutation, including indels.

[2] One-hot encoding is a technique that represents categorical variables as binary vectors. Each unique category is mapped to a specific position in the binary vector, where a value of one indicates its presence and a value of 0 indicates its absence.

Cao *et al.* (50) developed a high-throughput strategy to design, screen, and optimize 5′ UTRs for enhanced protein expression. By training a random forest (RF) model on a dataset of naturally occurring 5′ UTRs with high translation efficiencies, the authors created a total of ∼12,000 100-bp 5′ UTR libraries containing 3586 synthetic sequences (generated using a genetic algorithm) and 8414 natural sequences. The library was subsequently screened using a recombinase-mediated integration strategy. The RF model was developed to predict TE based on 5′ UTR characteristics, such as k-mer frequency, RNA folding energy, 5′ UTR length, and number of ORFs. Further validation of the top-predicted 5′ UTR hits that increased green fluorescent protein (GFP) production in HEK 293T cells enabled the authors to identify three synthetic 5′ UTRs that outperformed commonly used non-viral gene therapy plasmids in expressing protein payloads. This combination of experimental and computational techniques provides a robust strategy for the systematic discovery and engineering of 5′ UTRs to enhance protein expression.

Naturally occurring mRNA sequences, unlike synthetic ones, contain differing coding sequences and 3′ UTRs that can also affect translation regulation (63–65). Thus, the sequence motifs learned by existing ML approaches trained on a single dataset (single-task models) may not generalize well to other datasets (48, 55, 56). Instead of relying on single-task models, Zheng *et al.* (55, 56) proposed MTtrans, a multi-task learning model that integrates information from multiple datasets. The key component within MTtrans is the shared CNN layers, where task-specific inputs are transformed to task-specific feature maps. The feature maps are subsequently fed into task-specific network layers to predict related experimental quantities for the corresponding tasks. By using the same CNN layers across different tasks, MTtrans is forced to learn the shared patterns from multiple experimental systems. This approach is based on the fundamental assumption that the 5′ UTR sequence features extracted by shared CNN layers that can robustly predict translation rate across multiple datasets are more likely to be actual underlying regulatory elements governing mRNA translation. The results show that MTtrans outperforms both Optimus 5-Prime and FramePool for MRL prediction on the same MPRA datasets with synthetic 5′ UTRs and naturally occurring endogenous human 5′ UTRs. Using fluorescence-activated cell sorting coupled with a deep sequencing (FACS-seq) experiment, the authors further validated the impact of most motifs identified by the learned shared CNN encoder. These results indicate that sequence features identified through multi-task learning are generalizable across different experimental systems, highlighting its strength in identifying evolutionarily conserved sequence motifs.

To facilitate the generation of 5′ UTR sequences, Barazandeh *et al.* (54) proposed UTRGAN, a generative adversarial network (GAN) coupled with an optimization procedure to generate 5′ UTR sequences with desired features such as high ribosome load and TE, while still mimicking various properties of natural UTR sequences. Both UTRGAN's generator and discriminator are based on convolutional neural networks, where the generator learns to generate plausible sequences from samples drawn from a latent space (low-dimensional space containing compressed representations of the one-hot encoded input sequences), while the discriminator learns to distinguish the decoy data from the true ones. These two models are trained adversarially together in a zero-sum game until the discriminator model is no longer able to differentiate generated data from real ones, meaning the generator model is generating plausible examples. The results showed that the UTRGAN-generated 5′ UTRs maintain similar 4-mer, GC content, predicted MRL, TE, and minimum free energy (MFE) distributions compared to natural ones. The optimization pipeline in UTRGAN employs an iterative procedure where, in each iteration, the generated 5′ UTRs are evaluated by off-the-shelf ML prediction models, such as FramePool (48) and MTtrans (55, 56), to guide the direction of updates in the latent space through a gradient ascent algorithm.

More recently, several RNA language models (LMs) have emerged and exhibited better performance on 5′ UTR function-related prediction tasks, examples include RNA-FM (52) and UTR-LM (57). The key component behind these LMs is pre-training on large sequence datasets, which enables the extraction of meaningful semantic representations from raw RNA sequences. Compared to multi-task learning employed in MTtrans, the self- or semi-supervised nature of pre-training allows LMs to take advantage of massive amounts of unlabeled RNA sequence data, avoid solely relying on labeled information, and extract transferable meaningful semantic representations of RNA sequences across various biological systems. RNA-FM, proposed by Chen *et al.* (52), is an RNA foundation model based on the bidirectional encoder representations from transformers (BERT) language model architecture. It is built upon 12 transformer-based bidirectional encoder blocks and trained on 23 million sequences from the RNAcentral database (67) in a self-supervised manner. With simple CNN and multilayer perceptron (MLP) based prediction modules, RNA-FM can benefit various downstream tasks, including RNA secondary structure prediction, RNA 3D modeling, RNA-protein interaction modeling, and mRNA gene expression regulation modeling. Specifically, when trained on the same synthetic dataset containing 76,319 distinct random 5′ UTRs with lengths spanning from 25 to 100 nucleotides (49), the RNA-FM-based model achieves better correlations in MRL predictions than Optimus 5-Prime on two test sets consisting of 7600 random and 7600 human 5′ UTRs with varying lengths, respectively. UTR-LM, proposed by Chu *et al.* (57), is also a transformer-based model pre-trained to extract representations from the raw sequences *via* nucleotide masking and reconstruction. Compared to the RNA-FM, the initial input feature utilized in the pre-training stage not only contains 1D sequence information but is also further augmented with supervised information including predicted secondary structure and predicted MFE. The pre-trained model was fine-tuned on a variety of downstream tasks, outperforming Optimus, FramPool, MTtrans, Cao-RF, and RNA-FM on MRL and TE prediction in

terms of Spearman's rank correlation coefficient. Additionally, the authors demonstrated UTR-LM's ability to design high-efficiency 5′ UTRs; specifically, 211 new 5′ UTRs with high predicted TE values were tested experimentally, and the results confirmed that the top-designed sequence achieved a 32.5% increase in protein production level relative to well-established 5′ UTRs optimized for therapeutics.

These recent studies demonstrate that ML models can effectively learn the complex sequence-function relationships of 5′ UTRs from large datasets, providing both insights into the regulatory code as well as enabling the rational design of optimized 5′ UTR sequences for applications requiring precise control of protein expression levels. However, it remains crucial to understand the structural characteristics of these designed or optimized sequences. This understanding ensures that specific 2D/3D structures are formed and/or conserved sequences are appropriately exposed for the proper function of desired regulatory elements. Given that 5′ UTRs from human genes average ∼218 nucleotides in length (38), experimentally determining the 3D structure of the entire sequence can be challenging. Nevertheless, other experimental techniques, such as chemical probing (68–71), can provide RNA secondary structural information at single-nucleotide resolution and guide 2D structure prediction (72). Additionally, various computational tools can also offer valuable insights for analyzing the 2D/3D structures of 5′ UTRs (73–76). Interestingly, a recent study indicates that the structure of mature mRNAs correlates significantly with the structure of the corresponding cotranscriptionally folded pre-mRNAs (77). Therefore, when using various computational models to predict 2D/3D structures for designed 5′ UTRs, incorporating additional kinetic folding or cotranscriptional folding information might be useful.

3′ UTR and poly(A) tail

While not as critical as the 5′ UTR in regulating mRNA expression, the 3′ UTR contains several key regulatory elements that can still significantly impact overall protein expression by influencing mRNA lifetime within the cell. For instance, AU-rich elements (AREs)—usually 50 to 150-nucleotide sequences containing multiple AUUUA motifs—promote mRNA decay through interactions with ARE-binding proteins (31). Another important element is the microRNA (miRNA) response elements (MREs). When bound by miRNAs, it can lead to translational repression, preventing the mRNA from being translated into protein (78–80). Of utmost importance is the poly(A) tail, which contains binding sites for PABPs. This crucial element not only protects mRNA from Ribonucleases (RNases) but also facilitates ribosome loading through eukaryotic initiation factors (eIFs)-mediated end-to-end interactions during mRNA circularization (44, 81, 82). Although many miRNA and mRNA targets remain unknown, computational and experimental tools can be employed to identify potential miRNA targets (83–87). In principle, negative regulators such as MREs are excluded when designing 3′ UTRs, and one can utilize certain human mRNAs with 3′ UTRs devoid of MREs. However, excluding MREs could interfere with other essential regulatory elements within the 3′ UTRs that may either promote or suppress protein translation. Additionally, available human mRNA 3′ UTRs lacking MREs may not be optimal for efficient protein translation. All these factors should be carefully considered and optimized.

The human β-globin (hGb) 3′ UTR is widely used in IVT mRNA design due to its high effectiveness in enhancing protein expression (88). Recent studies have shown that other naturally occurring 3′ UTR sequences can match or even surpass the performance of the hGb 3′ UTR in this regard (89, 90). Currently, experimental approaches such as the systematic evolution of ligands by exponential enrichment (SELEX) remain crucial for optimizing 3′ UTR sequences. SELEX experiments have successfully identified 3′ UTR sequences that can significantly improve protein expression (91), with some enhancing total protein expression by over threefold across various coded proteins. Unlike the extensive computational efforts in 5′ UTRs design and optimization, 3′ UTRs have not yet benefited from computational modeling due to the lack of a comprehensive database for training and validation. Nevertheless, various structure prediction models can still provide valuable insights into 3′ UTR structures, helping to estimate their interactions with various 3′ UTR-binding factors (*e.g.*, ARE-binding proteins and MRE-binding miRNAs).

Optimizing ORF codon to improve translation efficiency

The ORF encodes the sequential information of a protein's amino acids. Due to codon degeneracy—where multiple distinct three-base pair codons can encode the same amino acid—one of the crucial aspects of ORF design involves finding a proper codon combination that matches the desired amino acid sequence and optimizes protein synthesis at the same time.

Codon optimization is the process of selecting an ideal codon for each amino acid. A common strategy aims to enhance translation rates by replacing rare codons with synonymous ones favored by the cellular environment, guided by the abundance of corresponding transfer RNAs (tRNAs) in host cells. This artificial synonymous codon usage bias can be quantified using the codon adaptation index (CAI) (92), a useful metric for estimating a gene's expression level. For a given ORF encoding a specific amino acid sequence, the CAI is calculated as the geometric mean of the relative frequencies of all codons used. The relative codon frequency of a particular codon is defined as its observed frequency divided by the frequency of its corresponding most prevalent synonymous codon within a reference set of highly expressed genes. Using the most prevalent codons for all amino acids in an ORF results in a CAI of 1.0. Notably, approved mRNA vaccines, such as BNT-162b2 and mRNA-1273, have optimized ORFs with CAI values exceeding 0.9 (23). Additionally, multiple experiments have demonstrated that codon selection can significantly influence the intracellular lifespan of mRNA (93–100). Therefore, the codon stabilization coefficient (CSC), derived

from the correlation between codon frequencies and mRNA half-lives, can serve as another quantitative metric for codon optimization (93).

The elongation of the amino acid chain requires the locally structured ORF sequence to partially unfold into a single-stranded form. This unfolding allows the ribosomal complex to move towards the 3′ end of the mRNA. However, the local structure of the ORF should not be so stable that it resists unfolding, as this could block amino acid chain elongation at that position (92). A model proposed by Paulo Gaspar *et al.* for optimizing mRNA ORF 2D structures employs a pseudo-minimum free energy (pseudo-MFE) based algorithm (101). This approach estimates the MFE of a given mRNA sequence by averaging the interaction energies of all possible single stem-loop conformations. Unlike methods that enumerate every possible 2D structure for a given sequence, this algorithm significantly reduces computational time, enabling the optimization of mRNA sequences exceeding 1000 codons. The goal of this model is to maximize the MFE for the sequence, thereby avoiding local stable structures that could potentially reduce translation efficiency.

Refining global sequence to improve structural stability

RNA, unlike DNA, is prone to base-catalyzed hydrolysis due to the 2′ hydroxyl group on its ribose sugar. This hydrolysis can occur spontaneously in solution, even without catalysts or enzymes, and is more likely in single-stranded regions where sensitive chemical groups are exposed (102). To maintain vaccine efficacy, mRNA vaccines must be stored at very low temperatures to minimize natural degradation (102). mRNA with well-structured regions, fewer single-stranded motifs, and higher GC content tends to have lower degradation rates, potentially allowing for more flexible storage requirements (103–105). The folded structures of ORFs can vary significantly based on codon selection, leading to substantial differences in structural stability among ORFs with different sequences. Therefore, a key principle in ORF design is prioritizing stable structures by identifying sequences that yield folded structures with lower MFE, thereby enhancing overall mRNA stability (104).

In the past, several computational tools have been developed to identify sequences capable of folding into low-energy structures at the 2D level. CDSfold, a model developed by Goro Terai *et al.*, aims to identify sequences with stable MFE structures (106). CDSfold searches for the ORF sequence that can be translated into the desired protein while simultaneously forming the most stable 2D structure by employing dynamic programming (107), a computationally efficient method to determine the MFE and its corresponding RNA structure. However, with the time complexity of $O(L^3)$ and space complexity of $O(L^2)$, where L represents the sequence length, the model may take several hours to analyze an mRNA ORF consisting of just a few thousand codons.

LinearDesign, a recently developed model, addresses the computational complexity inherent in sequence optimization tasks by employing a classical lattice parsing approach from computational linguistics (108). This innovative method draws an analogy between identifying the optimal mRNA sequence and selecting the most coherent sentence among similar-sounding alternatives in linguistics. By leveraging an enhanced left-to-right dynamic programming algorithm within the framework of lattice parsing, LinearDesign significantly reduces the time required to identify the mRNA sequence folding to the stable MFE structure (109). This advancement cuts processing time from several hours to minutes, marking a substantial improvement in computational efficiency. When tested for designing mRNA vaccines for SARS-CoV-2, LinearDesign successfully identified mRNA sequences capable of achieving over threefold higher protein expression in cells compared to established vaccines like BNT162b2 and mRNA-1273 (108). Additionally, it significantly increased antibody titers by up to 128 times in mice compared to the codon-optimization benchmark for mRNA vaccines targeting SARS-CoV-2 and varicella-zoster virus (VZV).

Predicting RNA structures from sequences is crucial for assessing and understanding structural stability. Most efforts in this field remain focused on 2D structures. For traditional (non-ML-based) RNA structure prediction methods, thermodynamic parameters and force fields are critical for accurately predicting both 2D and 3D RNA structures. For instance, in RNA 2D structure prediction, the nearest-neighbor free-energy model and Turner parameters are widely used (110, 111), and employed by models such as RNAfold (112), RNAstructure (72), mfold (113), and Vfold2D (114, 115, 116). Turner parameters represent a set of nearest-neighbor thermodynamic parameters for RNA folding, derived from experimental measurements. These parameters provide free energy and enthalpy/entropy changes for forming various RNA secondary structure motifs and are widely used in thermodynamics-based RNA secondary structure predictions. Existing thermodynamic-based RNA 2D structure prediction models face challenges for RNAs containing complex structural elements such as multibranched junction loops and pseudoknots (73). These limitations stem from the lack of accurate thermodynamic parameters for these elements. In practice, it is often useful to consider the consensus base pairs predicted from different models.

In addition to traditional folding energy, an alternative metric for assessing mRNA structural stability is the average unpaired probability (AUP), which reflects the overall 'unstructuredness' of the RNA (104). In principle, AUP may provide more relevant information about mRNA in-solution lifetime at a given storage temperature, because mRNA degradation may correlate with the total number of unpaired nucleotides (104). Wayment-Steele *et al.* developed the RiboTree model to optimize mRNA ORFs to achieve the lowest AUP (104). RiboTree utilizes a Monte Carlo tree search algorithm for stochastic minimization and the LinearPartition algorithm to compute the AUP for designed mRNA sequences (104, 109). The model successfully identified ORF sequences that can enhance the in-solution lifetime of COVID-19 mRNA by more than twofold in prediction.

To theoretically assess mRNA in-solution stability and minimize hydrolytic degradation of designed mRNA sequences, Leppek *et al.* proposed a linear regression model called DegScore (22) that can estimate mRNA in-solution degradation rate. Trained on a large-scale in-line chemical probing dataset comprising 3030 RNA fragments, DegScore can quantitatively capture properties related to mRNA degradation and predict mRNA in-solution lifetime (22). While ORF sequences corresponding to the lowest AUP, DegScore, or MFE are generally different, these metrics have been experimentally validated as useful for enhancing mRNA vaccine stability, both in solution and within cells (22).

ML-based models for RNA structure prediction have recently emerged, demonstrating significant capabilities in predicting 2D or 3D structures from sequences (73, 117). Several models have been developed and evaluated for designing short ORF sequences (102–130 nucleotides) in competitions like OpenVaccine (https://eternagame.org/challenges/10845741). By leveraging various architectures, such as autoencoders, CNNs, graph neural networks (GNNs), gated recurrent units (GRUs), and long short-term memory (LSTM) networks, these models have enhanced the prediction of mRNA degradation (118–123). However, ML-based models also face some limitations. They have not yet surpassed traditional RNA structure prediction models in terms of prediction accuracy, particularly for artificially designed RNAs, as demonstrated in the 15th critical assessment of structure prediction (CASP15) competition (73, 75). Additionally, current ML methods have not yet been extensively evaluated for large RNAs, especially for mRNAs containing thousands of codons.

Fine-tuning local sequence to extend the in-cell lifetime and increase expression fidelity

In addition to the global structure of ORFs, local structural elements such as loops, hairpins, junctions, and pseudoknots can significantly influence mRNA in-cell lifetime (124), local translation rate, and translation fidelity (125). These RNA local structures, formed within short sequence regions, play a crucial role in regulating mRNA in-cell degradation (124), ribosome moving rate (92), and potential translational frameshift stimulation (126).

mRNA can be quite unstable in cells due to the RNases-dependent RNA degradation (102, 127–129). The cellular RNase system includes two main types of enzymes: exoribonucleases (130–132)—enzymes that degrade RNA by removing terminal nucleotides from either the 5′ ends or the 3′ ends of the RNA molecule, and endoribonucleases (133, 134)—enzymes that cleave either single-stranded or double-stranded RNA chain by recognizing specific RNA sequence and structural motif (102, 129). To suppress endoribonuclease activation, several specific RNA sequences or structure elements should be avoided. Since single-stranded RNA (*i.e.* loops) is a common structural element recognized by various types of endoribonucleases (102, 129), designing mRNA sequences with fewer single-stranded regions is a practical way to increase mRNA in-cell lifetime. This goal aligns well with the optimization objective (lowest AUP) employed in the RiboTree model (104).

An interesting finding from the benchmark test of LinearDesign is that in-solution structural stability, indicated by the folding free energy, shows only a weak correlation with in-cell lifetime and protein expression for a typical mRNA sequence (108). This may be attributed to the optimization of the mRNA ORF for higher stability, which can hinder decoding progression due to ribosomal stalling caused by rare codon usage or the formation of stable local secondary structures (135–137). Ribosomal stalling during mRNA translation can result in ribosome collisions at specific sites, ultimately leading to mRNA cleavage and degradation (138–140). Such nonsense-mediated RNA decay significantly shortens the in-cell lifetime of mRNA and reduces protein expression efficiency (138–141). These mechanisms suggest the complexity of the secondary structure and folding stability-based ORF design for mRNA vaccines.

Protein cotranslational folding (142, 143), similar to RNA cotranscriptional folding, is a kinetic process highly sensitive to the elongation rate of the amino acid chain (142). Wild-type mRNA for an antigen typically incorporates specific codon sequences and structural elements within the ORF to regulate the elongation rate, ensuring the nascent amino acid chain folds into functional structures (142, 144, 145). In artificially designed mRNA for vaccines, however, the ORF sequence may be altered due to codon optimization. Consequently, the programmed elongation rate of the amino acid chain during translation may differ (125), potentially leading to the production of misfolded proteins (146). Structural misfolding can result in significant differences between the expressed protein and the target antigen. In more severe cases, such structural alterations may render the protein toxic to cells (*e.g.*, misfolded Aβ protein associated with Alzheimer's disease (147)). Several computational tools are available to predict protein cotranslational folding while accounting for translation rates (143, 146). These models can aid in codon optimization by identifying and avoiding sequences likely to result in misfolded structures.

Programmed translational frameshift, observed in many viruses and eukaryotes, requires specific RNA sequences and locally folded structures such as stem-loops or pseudoknots (126). Designed ORF sequences may inadvertently contain these sequence/structural elements, potentially stimulating unintended translational frameshifts (148, 149). This can lead to the expression of alternative proteins, increasing the risk of side effects. COVID-19 mRNA vaccines have been reported to produce undesired proteins in clinical applications (150). Studies have also shown that up to 30% of expressed proteins in host cells have altered amino acid sequences due to translational frameshifts (150, 151).

However, current ORF sequence design tools cannot explicitly consider local structures. One potential approach is to predict global structures using various state-of-the-art models, including physics-based and machine learning-based models, and then search for consensus local structural elements within the predictions. Table 1 provides a summary of selected computational tools for RNA structure prediction. An

Table 1
RNA structure prediction tools

2D structure prediction		3D structure prediction	
Model	**Method**	**Model**	**Method**
RNAstructure (72)	TM	iFold (184)	MD
RNAfold (112)	TM	SimRNA (185)	MC
RNAalifold (186)	TM	RNAComposer (187)	AS
PKNOTS (188)	TM	FARNA (189)	MC
Mfold (113)	TM	MC-Sym (190)	AS
Vfold2D (114, 115, 116)	TM	FARFAR2 (191)	AS
HotKnots (192, 193)	TM	ARES (194)	ML
CentroidFold (195)	TM	IsRNA (196, 197)	CGMD
IPknot (198)	TM	RNAJP (199)	MD/MC
LinearFold (200)	TM	RNA3DCNN (201)	ML
TurboFold (202)	TM	PaxNet (203)	ML
SPOT-RNA (204)	ML	DeepFoldRNA (205)	ML
E2Efold (206)	ML	trRosettaRNA (207)	ML
MXfold2 (208)	ML	epRNA (209)	ML
EternaFold (210)	ML	E2Efold-3D (211)	ML
Ufold (212)	ML	ReseTTAFoldNA (213)	ML
DMfold (214)	ML	DRFold (215)	ML
CNNfold (216)	ML	AlphaFold3 (217)	ML
RNAformer (218)	ML	3dRNA (219)	AS
2dRNA (220)	ML	Vfold-Pipeline (221)	AS

AS, Assembly; CGMD, Coarse-Grained Molecule Dynamic; MC, Monte Carlo; MD, Molecule Dynamic; ML, Machine Learning; TM, Thermodynamic.

alternative approach involves computing base pairing probabilities from the partition function and subsequently inferring local structures from these probabilities (152, 153). Identifying local structures can help pinpoint regions susceptible to problematic conformations, which may adversely impact translation efficiency and fidelity.

Multi-objective optimization

The sequence design of mRNA vaccines generally involves multiple objectives. Given the intrinsic properties of mRNA molecules, achieving all objectives within a single sequence can be challenging. For example, maximizing structural stability often conflicts with the optimal codon selection.

Recent studies on ORF design have demonstrated that optimal antigen expression cannot be achieved by simply optimizing CAI, MFE, or AUP alone. Instead, the best-performing ORF sequences typically find a balance between CAI and structural stability (22). Based on this finding, several approaches have been proposed to optimize both translation efficiency and mRNA stability within the cell, thereby achieving optimal protein expression efficiency for designed vaccines. These models utilize a combined scoring function that integrates CAI with MFE or AUP. For example, CDSfold uses $MFE \times CAI^{\lambda}$ to combine the two factors, while LinearDesign uses $MFE - \lambda \log(CAI)$. However, determining the optimal parameter λ in these scores is nontrivial, as benchmark experiments have shown that the best-performing ORF sequences correspond to different λ values for different protein targets of SARS-CoV-2 and VZV (108). Rather than relying on a single value, LinearDesign employs a range of λ to explore a broader region of candidate sequences and evaluates all candidates through additional experiments (108). This approach enables exploring previously unreachable yet important sequence space.

As shown in Figure 3 for the computationally designed mRNA vaccines for SARS-CoV-2 and VZV, identifying the optimal ORF sequence requires balancing multiple factors beyond CAI and structural stability. Codon selection, local structural elements, and various other factors must be considered concurrently. Although LinearDesign can identify potential high-efficiency regions for mRNA sequence design

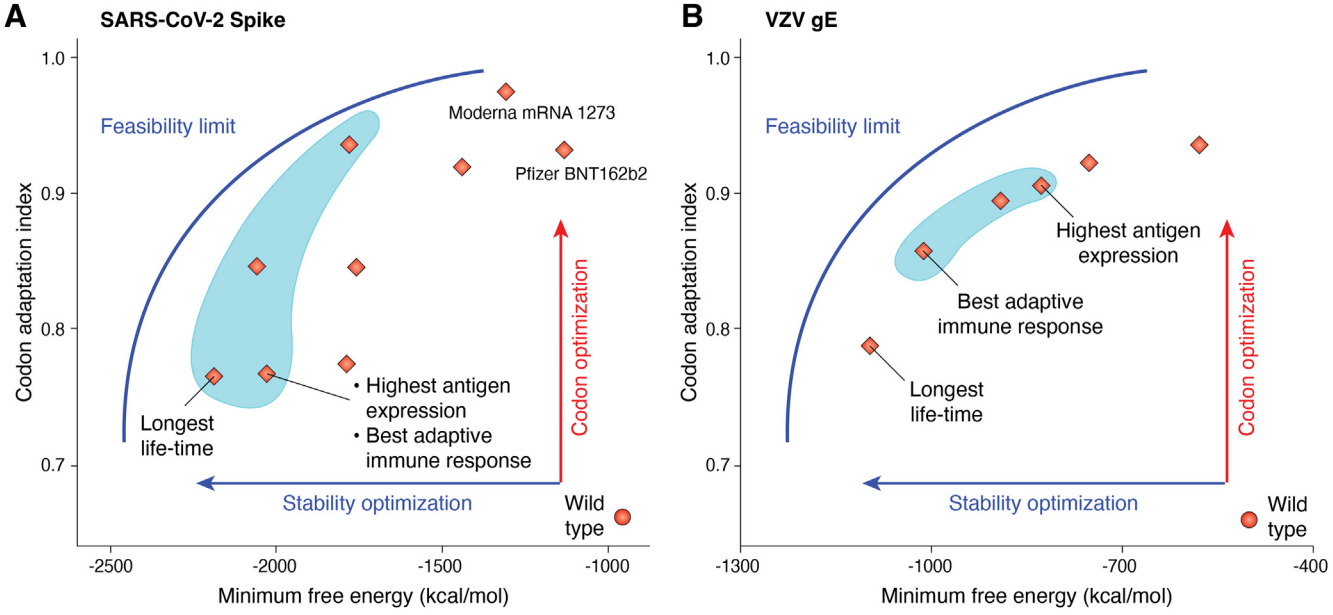

Figure 3. Computationally designed mRNA vaccines. Sequences for (*A*) severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and (*B*) varicella zoster virus (VZV), together with their calculated codon adaptation indices (CAIs), minimum free energies (MFEs), and various experimentally measured properties. In both (*A*) and (*B*), the areas to the right of the feasibility limit curves represent possible regions for designing mRNA sequences. Light blue shaded areas indicate potential high-efficiency mRNA vaccine regions predicted by LinearDesign (108). For mRNA vaccines that encode either SARS-CoV-2 spike protein or VZV gE protein, the best-performing one (*i.e.* inducing the strongest adaptive immune response) has neither the highest CAI nor the lowest MFE. Data is obtained from reference (108).

(see Fig. 3), determining the exact "best" sequence among the suggested candidates remains a challenge. The integration of different mRNA components (*e.g.*, UTRs and ORF) introduces additional challenges, such as long-range interactions between ORFs and UTRs, and effects induced by trans-activating regulators within the UTRs.

Recent advancements in deep learning-based techniques have demonstrated significant advantages in multi-objective optimizations (154, 155). As discussed in the section *Optimizing and designing UTR sequence to enhance ribosome loading and translation efficiency*, several approaches have been developed to optimize 5′ UTRs by considering multiple objectives. Nevertheless, applying deep learning methods to ORF design remains challenging due to the scarcity of existing databases. To build comprehensive databases, mRNA ribosome load, in-solution stability, in-cell lifetime, and translation efficiency need to be experimentally measured across a broad range of diverse sequences. Addressing this need, Leppek *et al.* developed PERSIST-seq, an RNA sequencing-based platform that systematically delineates key properties of designed mRNAs (22). Currently, 233 mRNA vaccine sequences, along with their associated key properties, have been deposited through the application of PERSIST-seq (22). Deep learning models developed to predict gene expression (156, 157) may be extended to IVT mRNA as available databases rapidly expand. Furthermore, given the remarkable capabilities of generative artificial intelligence (GAI) in producing various types of biological data (158–160), transformer-based deep neural networks, such as those used in large language models (LLMs), show great potential for designing novel mRNA vaccine sequences.

Utilizing RNA modifications to improve mRNA vaccine efficacy

Hundreds of natural modifications have been identified and characterized for RNAs. These modifications play a critical role in regulating mRNA functions (161–166), ranging from reducing mRNA immunogenicity (163, 164) and enhancing mRNA stability (162) to modulating mRNA translation (163, 164). RNA modifications have been employed in vaccine design and clinical and preclinical mRNA therapeutic applications (164). A notable application is in SARS-CoV-2 mRNA vaccines, such as Moderna's mRNA-1273 and BioNTech/Pfizer's BNT-162b2, where the modified nucleobase N1-methyl pseudouridine (m1Ψ) has proven crucial for vaccine efficacy (163, 167). Another widely utilized modification in mRNA therapeutics is pseudouridine (Ψ), which can significantly enhance mRNA translation efficiency (168, 169). Beyond SARS-CoV-2 mRNA vaccines, other examples include vaccines targeting influenza viruses (170), cytomegalovirus (CMV) (171), HIV (172), Ebola (173), Zika (174), and human metapneumovirus (hMPV) (175), *etc.* Here, we focus on the impact of nucleoside modifications in 5′ UTR, ORF, and 3′ UTR regions on mRNA vaccine efficacy from a structural perspective. For a comprehensive overview of how other modifications—such as those in 5′-cap, backbone, and poly(A) tail modifications, as well as the enzymatic ligation of nuclease-resistant oligonucleotides to the poly(A) tail (176)—affect mRNA efficacy, readers are directed to additional reviews (164, 165).

From a structural perspective, RNA modifications can influence RNA folding by altering the structural and energetic properties of base pairing and stacking, including conformational flexibility, groove hydrophobicity, and the stability of long-range contacts (177–179). As a result, these modified bases could potentially alter RNA secondary structure. For instance, a chemical probing study suggested that RNAs containing m1Ψ and uridine may adopt distinct secondary structures (128). Additionally, optical melting experiments on synthetic short RNA duplexes indicated that Ψ and m1Ψ modifications, compared with the unmodified uridine (U), could increase the stability by 0.25 and 0.18 kcal/mol, respectively, for each base pair (128). However, studies have also indicated that m1Ψ can reduce translational fidelity by causing +1 ribosomal frameshifting, producing altered proteins (150, 151). Therefore, various aspects must be considered when designing predictive computational models.

To predict the impact of modified nucleotides in RNA structure and stability, the ViennaRNA package includes six modified nucleotides such as Ψ by incorporating energy corrections derived from experimental data (180, 181), and RNAstructure employs linear regression to fit thermodynamic parameters for modified nucleotides such as N6-methyladenosine (m6A) (182). Integrating modified bases in RNA structure prediction remains an ongoing endeavor. With the expanding experimental data on the thermodynamics of modified nucleotides, we anticipate that more accurate models will emerge to facilitate the design and optimization of mRNA vaccine sequences and structures.

Summary and perspective

The rapid advancement of mRNA vaccine technology has provided an increasingly promising tool against infectious diseases and cancers. While mRNA vaccines have demonstrated significant success in preventing COVID-19, their application in other areas remains limited. Various computational models, including advanced ML approaches trained on large experimental 5′ UTR datasets, now enable the evaluation, optimization, and rational design of mRNA sequences for diverse applications. However, the scarcity of data poses a major obstacle in applyingpowerful ML models to mRNA vaccine design. Additionally, the limited accuracy and scalability of the RNA 2D/3D structure prediction tools —especially for sequences containing thousands of nucleotides—further hinder the design process. Looking forward, several areas warrant further research and development.

1. Integrated sequence design/optimization: while current approaches focus on optimizing individual mRNA components, future research should aim to develop integrated models that consider both the global mRNA structure which determines in-solution stability and the local structure/sequence motif which affects the mRNA translation efficiency, translation fidelity, and in-cell degradation rate.

These models should also consider alternative folding pathways, such as cotranscriptional folding, and intermediate (subpopulated) states, which can lead to structures different from the thermodynamic equilibrium MFE structure.

2. Expanding current mRNA vaccine-related databases: The expansion will likely require extensive experimental measurements of various properties for mRNA sequences, such as translation efficiency, translation fidelity, in-solution stability, in-cell lifetime, and reactogenicity. Additionally, experimental thermodynamic parameters for modified nucleotides can further benefit various RNA 2D/3D structure prediction tools (see Table 1).
3. Balancing multiple objectives: Current mRNA vaccine design principles often focus on individual objectives. Future research should prioritize developing optimization algorithms that can effectively balance competing design goals to identify a globally optimal mRNA sequence. Achieving this will require a deep understanding of the underlying mechanisms driving the immune response triggered by mRNA vaccines.
4. Experimental validation: while computational approaches offer powerful tools for mRNA vaccine design, rigorous experimental validation remains essential. Developing high-throughput screening methods to test computationally designed sequences will be crucial for advancing the field.

In conclusion, computational approaches hold tremendous potential for advancing mRNA vaccine design and optimization. By integrating biological knowledge, machine learning techniques, and experimental validation, these methods can greatly improve the efficacy, stability, and safety of mRNA vaccines for emerging infectious diseases and therapeutic areas such as cancer immunotherapy.

Data availability

The data supporting the findings of this study are available in the manuscript.

Declaration of generative AI in scientific writing

The large language model (LLM), Claude, developed by Anthropic was used for language refinement only.

Author contributions—S. Z., L.J., and Y. Z. writing–original draft, S. Z., L.J., and Y. Z. investigation; S. C. writing–review & editing; S. C. supervision; S. C. funding acquisition; S. C. conceptualization.

Funding and additional information—This work was supported by the National Institutes of Health under Grants R35-GM134919 and U54 AI170660 and the National Science Foundation under Grant Number CHE 2154924. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of interest—The authors declare that they have no conflicts of interest with the contents of this article.

Abbreviations—The abbreviations used are: ARE, AU-rich element; AUP, average unpaired probability; BERT, bidirectional encoder representations from transformers; CAI, codon adaptation index; CDS, coding sequence; CNN, convolutional neural networks; COVID-19, coronavirus disease 2019; eGFP, enhanced green fluorescent protein; eIFs, eukaryotic initiation factors; FACS-seq, fluorescence-activated cell sorting coupled with deep sequencing; GAN, generative adversarial networks; GFP, green fluorescent protein; hGb, human β-globin; IRE, iron response element; IVT, *in vitro* transcription; LM, language model; LNP, lipid nanoparticle; LSTM, long short-term memory; m^7G, 7-methylguanosine; MFE, minimum free energy; miRNA, microRNA; ML, machine learning; MLP, multilayer perceptron; MPRA, massively parallel reporter assay; MRE, microRNA response element; MRL, mean ribosome load; mRNA, messenger RNA; ORF, open reading frame; PABP, poly(A)-binding protein; PCR, polymerase chain reaction; pre-mRNA, precursor mRNA; pseudo-MFE, pseudo-minimum free energy; PTR, protein-to-mRNA ratios; RBS, ribosome binding site; RF, random forests; RNase, Ribonuclease; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; SNV, single nucleotide variant; TE, translation efficiency; tRNA, transfer RNA; uORF, upstream open reading frame; UTR, untranslated region; VZV, varicella zoster virus.

References

1. Shattock, A. J., Johnson, H. C., Sim, S. Y., Carter, A., Lambach, P., Hutubessy, R. C., *et al.* (2024) Contribution of vaccination to improved survival and health: modelling 50 years of the Expanded Programme on Immunization. *Lancet* **403**, 2307–2316
2. Plotkin, S. A. (2005) Vaccines: past, present and future. *Nat. Med.* **11**, S5–S11
3. Ada, G. (2001) Vaccines and vaccination. *N. Engl. J. Med.* **345**, 1042–1053
4. Pardi, N., Hogan, M. J., Porter, F. W., and Weissman, D. (2018) mRNA vaccines—a new era in vaccinology. *Nat. Rev. Drug Discov.* **17**, 261–279
5. Saxena, M., van der Burg, S. H., Melief, C. J., and Bhardwaj, N. (2021) Therapeutic cancer vaccines. *Nat. Rev. Cancer* **21**, 360–378
6. Bonilla, F. A., and Oettgen, H. C. (2010) Adaptive immunity. *J. Allergy Clin. Immunol.* **125**, S33–S40
7. Cooper, M. D., and Alder, M. N. (2006) The evolution of adaptive immune systems. *Cell* **124**, 815–822
8. Chaudhary, N., Weissman, D., and Whitehead, K. A. (2021) mRNA vaccines for infectious diseases: principles, delivery and clinical translation. *Nat. Rev. Drug Discov.* **20**, 817–838
9. Corbett, K. S., Edwards, D. K., Leist, S. R., Abiona, O. M., Boyoglu-Barnum, S., Gillespie, R. A., *et al.* (2020) SARS-CoV-2 mRNA vaccine design enabled by prototype pathogen preparedness. *Nature* **586**, 567–571
10. Pollard, A. J., and Bijker, E. M. (2021) A guide to vaccinology: from basic principles to new developments. *Nat. Rev. Immunol.* **21**, 83–100
11. Rosa, S. S., Prazeres, D. M., Azevedo, A. M., and Marques, M. P. (2021) mRNA vaccines manufacturing: challenges and bottlenecks. *Vaccine* **39**, 2190–2200
12. Freyn, A. W., da Silva, J. R., Rosado, V. C., Bliss, C. M., Pine, M., Mui, B. L., *et al.* (2020) A multi-targeting, nucleoside-modified mRNA influenza virus vaccine provides broad protection in mice. *Mol. Ther.* **28**, 1569–1584
13. Crommelin, D. J., Anchordoquy, T. J., Volkin, D. B., Jiskoot, W., and Mastrobattista, E. (2021) Addressing the cold reality of mRNA vaccine stability. *J. Pharm. Sci.* **110**, 997–1001
14. Pardi, N., Hogan, M. J., and Weissman, D. (2020) Recent advances in mRNA vaccine technology. *Curr. Opin. Immunol.* **65**, 14–20
15. Al Fayez, N., Nassar, M. S., Alshehri, A. A., Alnefaie, M. K., Almughem, F. A., Alshehri, B. Y., *et al.* (2023) Recent advancement in mRNA vaccine development and applications. *Pharmaceutics* **15**, 1972

16. Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., *et al.* (2021) Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *N. Engl. J. Med.* **384**, 403–416
17. Teo, S. P. (2022) Review of COVID-19 mRNA vaccines: BNT162b2 and mRNA-1273. *J. Pharm. Pract.* **35**, 947–951
18. Shabu, A., and Nishtala, P. S. (2023) Safety outcomes associated with the moderna COVID-19 vaccine (mRNA-1273): a literature review. *Expert Rev. Vaccin.* **22**, 393–409
19. Xu, W., Ren, W., Wu, T., Wang, Q., Luo, M., Yi, Y., *et al.* (2023) Real-world safety of COVID-19 mRNA vaccines: a systematic review and meta-analysis. *Vaccines* **11**, 1118
20. Laurini, G. S., Montanaro, N., Broccoli, M., Bonaldo, G., and Motola, D. (2023) Real-life safety profile of mRNA vaccines for COVID-19: an analysis of VAERS database. *Vaccine* **41**, 2879–2886
21. Kim, Y. A., Mousavi, K., Yazdi, A., Zwierzyna, M., Cardinali, M., Fox, D., *et al.* (2024) Computational design of mRNA vaccines. *Vaccine* **42**, 1831–1840
22. Leppek, K., Byeon, G. W., Kladwang, W., Wayment-Steele, H. K., Kerr, C. H., Xu, A. F., *et al.* (2022) Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. *Nat. Commun.* **13**, 1536
23. Xia, X. (2021) Detailed dissection and critical evaluation of the Pfizer/BioNTech and Moderna mRNA vaccines. *Vaccines* **9**, 734
24. Beckert, B., and Masquida, B. (2011) Synthesis of RNA by in vitro transcription. In: Nielsen, H., ed. *RNA: Methods and Protocols*, 41, Hunana Press, Totowa, NJ: 29
25. Ramanathan, A., Robb, G. B., and Chan, S. H. (2016) mRNA capping: biological functions and applications. *Nucleic Acids Res.* **44**, 7511–7526
26. Furuichi, Y. (2015) Discovery of m7G-cap in eukaryotic mRNAs. *Proc. Jpn. Acad. Ser. B.* **91**, 394–409
27. Pan, R., Kindler, E., Cao, L., Zhou, Y., Zhang, Z., Liu, Q., *et al.* (2022) N7-Methylation of the coronavirus RNA cap is required for maximal virulence by preventing innate immune recognition. *MBio* **13**, e03662-21
28. Laursen, B. S., Sørensen, H. P., Mortensen, K. K., and Sperling-Petersen, H. U. (2005) Initiation of protein synthesis in bacteria. *Microbiol. Mol. Biol. Rev.* **69**, 101–123
29. Barrett, L. W., Fletcher, S., and Wilton, S. D. (2012) Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cell. Mol. Life Sci.* **69**, 3613–3634
30. Thompson, S. R. (2012) Tricks an IRES uses to enslave ribosomes. *Trends Microbiol.* **20**, 558–566
31. Kozak, M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361**, 13–37
32. Mudge, J. M., Ruiz-Orera, J., Prensner, J. R., Brunet, M. A., Calvet, F., Jungreis, I., *et al.* (2022) Standardized annotation of translated open reading frames. *Nat. Biotechnol.* **40**, 994–999
33. Sieber, P., Platzer, M., and Schuster, S. (2018) The definition of open reading frame revisited. *Trends Genet.* **34**, 167–170
34. Lee, Y., and Rio, D. C. (2015) Mechanisms and regulation of alternative pre-mRNA splicing. *Annu. Rev. Biochem.* **84**, 291–323
35. Pichon, X., A Wilson, L., Stoneley, M., Bastide, A., A King, H., Somers, J., *et al.* (2012) RNA binding protein/RNA element interactions and the control of translation. *Curr. Protein Pept. Sci.* **13**, 294–304
36. Guhaniyogi, J., and Brewer, G. (2001) Regulation of mRNA stability in mammalian cells. *Gene* **265**, 11–23
37. Hinnebusch, A. G., Ivanov, I. P., and Sonenberg, N. (2016) Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science* **352**, 1413–1416
38. Leppek, K., Das, R., and Barna, M. (2018) Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat. Rev. Mol. Cell Biol.* **19**, 158–174
39. Cohen-Chalamish, S., Hasson, A., Weinberg, D., Namer, L. S., Banai, Y., Osman, F., *et al.* (2009) Dynamic refolding of IFN-γ mRNA enables it to function as PKR activator and translation template. *Nat. Chem. Biol.* **5**, 896–903
40. Dethoff, E. A., and Weeks, K. M. (2019) Effects of refolding on large-scale RNA structure. *Biochemistry* **58**, 3069–3077
41. Soemedi, R., Cygan, K. J., Rhine, C. L., Glidden, D. T., Taggart, A. J., Lin, C. L., *et al.* (2017) The effects of structure on pre-mRNA processing and stability. *Methods* **125**, 36–44
42. Morris, D. R., and Geballe, A. P. (2000) Upstream open reading frames as regulators of mRNA translation. *Mol. Cell. Biol.* **20**, 8635–8642
43. Muckenthaler, M. U., Galy, B., and Hentze, M. W. (2008) Systemic iron homeostasis and the iron-responsive element/iron-regulatory protein (IRE/IRP) regulatory network. *Annu. Rev. Nutr.* **28**, 197–213
44. Gingold, H., and Pilpel, Y. (2011) Determinants of translation efficiency and accuracy. *Mol. Syst. Biol.* **7**, 481
45. Walden, W. E., Selezneva, A. I., Dupuy, J., Volbeda, A., Fontecilla-Camps, J. C., Theil, E. C., *et al.* (2006) Structure of dual function iron regulatory protein 1 complexed with ferritin IRE-RNA. *Science* **314**, 1903–1908
46. Huang, Y. C., Tse, S. W., Iacovelli, J., McKinney, K., and Valiante, N. (2019) *Immunomodulatory Therapeutic mRNA Compositions Encoding Activating Oncogene Mutation Peptides*, U.S. Patent and Trademark Office, Alexandria, VA (U.S. Patent No. US-2019175727-A1)
47. Frankish, A., Carbonell-Sala, S., Diekhans, M., Jungreis, I., Loveland, J. E., Mudge, J. M., *et al.* (2023) GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **51**, D942–D949
48. Karollus, A., Avsec, Ž., and Gagneur, J. (2021) Predicting mean ribosome load for 5′UTR of any length using deep learning. *PLoS Comput. Biol.* **17**, e1008982
49. Sample, P. J., Wang, B., Reid, D. W., Presnyak, V., McFadyen, I. J., Morris, D. R., *et al.* (2019) Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nat. Biotechnol.* **37**, 803–809
50. Cao, J., Novoa, E. M., Zhang, Z., Chen, W. C., Liu, D., Choi, G. C., *et al.* (2021) High-throughput 5′ UTR engineering for enhanced protein production in non-viral gene therapies. *Nat. Commun.* **12**, 4138
51. Akiyama, M., and Sakakibara, Y. (2022) Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. *NAR: Genomics Bioinf.* **4**, lqac012
52. [preprint] Chen, J., Hu, Z., Sun, S., Tan, Q., Wang, Y., Yu, Q., *et al.* (2022) Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv*. https://doi.org/10.48550/arXiv.2204.00300
53. Castillo-Hair, S. M., and Seelig, G. (2021) Machine learning for designing next-generation mRNA therapeutics. *Acc. Chem. Res.* **55**, 24–34
54. [preprint] Barazandeh, S., Ozden, F., Hincer, A., Seker, U. O. S., and Cicek, A. E. (2023) Utrgan: learning to generate 5'utr sequences for optimized translation efficiency and gene expression. *bioRxiv*. https://doi.org/10.1101/2023.01.30.526198
55. Zheng, W., Fong, J. H., Wan, Y. K., Chu, A. H., Huang, Y., Wong, A. S., *et al.* (2023) Translation rate prediction and regulatory motif discovery with multi-task learning. In: Tang, H., ed. *International Conference on Research in Computational Molecular Biology*, 154, Springer, Cham: 139
56. Zheng, W., Fong, J. H., Wan, Y. K., Chu, A. H., Huang, Y., Wong, A. S., *et al.* (2023) Discovery of regulatory motifs in 5′ untranslated regions using interpretable multi-task learning models. *Cell Syst* **14**, 1103–1112
57. Chu, Y., Yu, D., Li, Y., Huang, K., Shen, Y., Cong, L., *et al.* (2024) A 5′ UTR language model for decoding untranslated regions of mRNA and function predictions. *Nat. Mach. Intell.* **6**, 449–460
58. Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., *et al.* (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868
59. Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., *et al.* (2014) Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587
60. Eichhorn, S. W., Guo, H., McGeary, S. E., Rodriguez-Mias, R. A., Shin, C., Baek, D., *et al.* (2014) mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. *Mol. Cell.* **56**, 104–115
61. Andreev, D. E., O'Connor, P. B., Fahey, C., Kenny, E. M., Terenin, I. M., Dmitriev, S. E., *et al.* (2015) Translation of 5′ leaders is pervasive in genes resistant to eIF2 repression. *Elife* **4**, e03971

62. Xiao, Y., Tong, H., Yang, X., Xu, S., Pan, Q., Qiao, F., *et al.* (2016) Genome-wide dissection of the maize ear genetic architecture using multiple populations. *New Phytol.* **210**, 1095–1106
63. Floor, S. N., and Doudna, J. A. (2016) Tunable protein synthesis by transcript isoforms in human cells. *elife* **5**, e10921
64. Eraslan, B., Wang, D., Gusic, M., Prokisch, H., Hallström, B. M., Uhlén, M., *et al.* (2019) Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. *Mol. Syst. Biol.* **15**, e8513
65. Vogel, C., de Sousa Abreu, R., Ko, D., Le, S. Y., Shapiro, B. A., Burns, S. C., *et al.* (2010) Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* **6**, 400
66. Avsec, Ž., Kreuzhuber, R., Israeli, J., Xu, N., Cheng, J., Shrikumar, A., *et al.* (2019) The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.* **37**, 592–600
67. Sweeney, B. A., Petrov, A. I., Ribas, C. E., Finn, R. D., Bateman, A., Szymanski, M., *et al.* (2021) RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res.* **49**, D212–D220
68. Wang, X. W., Liu, C. X., Chen, L. L., and Zhang, Q. C. (2021) RNA structure probing uncovers RNA structure-dependent biological functions. *Nat. Chem. Biol.* **17**, 755–766
69. Spitale, R. C., and Incarnato, D. (2023) Probing the dynamic RNA structurome and its functions. *Nat. Rev. Genet.* **24**, 178–196
70. Li, B., Cao, Y., Westhof, E., and Miao, Z. (2020) Advances in RNA 3D structure modeling using experimental data. *Front. Genet.* **11**, 574485
71. Cao, X., Zhang, Y., Ding, Y., and Wan, Y. (2024) Identification of RNA structures and their roles in RNA functions. *Nat. Rev. Mol. Cell Biol.* **25**, 784–801
72. Reuter, J. S., and Mathews, D. H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinf.* **11**, 1–9
73. Das, R., Kretsch, R. C., Simpkin, A. J., Mulvaney, T., Pham, P., Rangan, R., *et al.* (2023) Assessment of three-dimensional RNA structure prediction in CASP15. *Proteins: Struct. Funct. Bioinf.* **91**, 1747–1770
74. Bernetti, M., and Bussi, G. (2023) Integrating experimental data with molecular simulations to investigate RNA structural dynamics. *Curr. Opin. Struct. Biol.* **78**, 102503
75. Li, J., Zhang, S., and Chen, S. J. (2023) Advancing RNA 3D structure prediction: exploring hierarchical and hybrid approaches in CASP15. *Proteins: Struct. Funct. Bioinf.* **91**, 1779–1789
76. Sato, K., and Hamada, M. (2023) Recent trends in RNA informatics: a review of machine learning and deep learning for RNA secondary structure prediction and RNA drug discovery. *Brief. Bioinf.* **24**, bbad186
77. Yu, G., Liu, Y., Li, Z., Deng, S., Wu, Z., Zhang, X., *et al.* (2023) Genome-wide probing of eukaryotic nascent RNA structure elucidates cotranscriptional folding and its antimutagenic effect. *Nat. Commun.* **14**, 5853
78. Wu, L., Fan, J., and Belasco, J. G. (2006) MicroRNAs direct rapid deadenylation of mRNA. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 4034–4039
79. Ghosh, T., Soni, K., Scaria, V., Halimani, M., Bhattacharjee, C., and Pillai, B. (2008) MicroRNA-mediated up-regulation of an alternatively polyadenylated variant of the mouse cytoplasmic β-actin gene. *Nucleic Acids Res.* **36**, 6318–6332
80. Wei, L., Li, S., Zhang, P., Hu, T., Zhang, M. Q., Xie, Z., *et al.* (2021) Characterizing microRNA-mediated modulation of gene expression noise and its effect on synthetic gene circuits. *Cell Rep.* **36**, 109573
81. Wells, S. E., Hillner, P. E., Vale, R. D., and Sachs, A. B. (1998) Circularization of mRNA by eukaryotic translation initiation factors. *Mol. Cell.* **2**, 135–140
82. Groft, C. M., and Burley, S. K. (2002) Recognition of eIF4G by rotavirus NSP3 reveals a basis for mRNA circularization. *Mol. Cell.* **9**, 1273–1283
83. Kiriakidou, M., Nelson, P. T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z., *et al.* (2004) A combined computational-experimental approach predicts human microRNA targets. *Genes Dev.* **18**, 1165–1178
84. John, B., Sander, C., and Marks, D. S. (2006) Prediction of human microRNA targets. In: Ying, S. Y., ed. *Methods in Molecular Biology*, 113, Humana Press, Totowa, NJ: 101
85. Riolo, G., Cantara, S., Marzocchi, C., and Ricci, C. (2020) miRNA targets: from prediction tools to experimental validation. *Methods Protoc.* **4**, 1
86. Chen, Y., and Wang, X. (2020) miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res.* **48**, D127–D131
87. Gu, T., Zhao, X., Barbazuk, W. B., and Lee, J. H. (2021) miTAR: a hybrid deep learning-based approach for predicting miRNA targets. *BMC bioinf.* **22**, 1–16
88. Zarghampoor, F., Azarpira, N., Khatami, S. R., Behzad-Behbahani, A., and Foroughmand, A. M. (2019) Improved translation efficiency of therapeutic mRNA. *Gene* **707**, 231–238
89. Siegel, D. A., Le Tonqueze, O., Biton, A., Zaitlen, N., and Erle, D. J. (2022) Massively parallel analysis of human 3′ UTRs reveals that AU-rich element length and registration predict mRNA destabilization. *G3: Genes, Genomes, Genet.* **12**, jkab404
90. Litterman, A. J., Kageyama, R., Le Tonqueze, O., Zhao, W., Gagnon, J. D., Goodarzi, H., *et al.* (2019) A massively parallel 3′ UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization. *Genome Res.* **29**, 896–906
91. von Niessen, A. G. O., Poleganov, M. A., Rechner, C., Plaschke, A., Kranz, L. M., Fesser, S., *et al.* (2019) Improving mRNA-based therapeutic gene delivery by expression-augmenting 3′ UTRs identified by cellular library screening. *Mol. Ther.* **27**, 824–836
92. Sharp, P. M., and Li, W. H. (1987) The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295
93. Presnyak, V., Alhusaini, N., Chen, Y. H., Martin, S., Morris, N., Kline, N., *et al.* (2015) Codon optimality is a major determinant of mRNA stability. *Cell* **160**, 1111–1124
94. Burow, D. A., Martin, S., Quail, J. F., Alhusaini, N., Coller, J., and Cleary, M. D. (2018) Attenuated codon optimality contributes to neural-specific mRNA decay in Drosophila. *Cell Rep.* **24**, 1704–1712
95. Bazzini, A. A., Del Viso, F., Moreno-Mateos, M. A., Johnstone, T. G., Vejnar, C. E., Qin, Y., *et al.* (2016) Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. *EMBO J.* **35**, 2087–2103
96. Harigaya, Y., and Parker, R. (2016) Codon optimality and mRNA decay. *Cell Res.* **26**, 1269–1270
97. Hia, F., Yang, S. F., Shichino, Y., Yoshinaga, M., Murakawa, Y., Vandenbon, A., *et al.* (2019) Codon bias confers stability to human mRNAs. *EMBO Rep.* **20**, e48220
98. Mishima, Y., and Tomari, Y. (2016) Codon usage and 3′ UTR length determine maternal mRNA stability in zebrafish. *Mol. Cell.* **61**, 874–885
99. de Freitas Nascimento, J., Kelly, S., Sunter, J., and Carrington, M. (2018) Codon choice directs constitutive mRNA levels in trypanosomes. *elife* **7**, e32467
100. Wu, Q., Medina, S. G., Kushawah, G., DeVore, M. L., Castellano, L. A., Hand, J. M., *et al.* (2019) Translation affects mRNA stability in a codon-dependent manner in human cells. *elife* **8**, e45396
101. Gaspar, P., Moura, G., Santos, M. A., and Oliveira, J. L. (2013) mRNA secondary structure optimization using a correlated stem–loop prediction. *Nucleic Acids Res.* **41**, e73
102. Tourriere, H., Chebli, K., and Tazi, J. (2002) mRNA degradation machines in eukaryotic cells. *Biochimie* **84**, 821–837
103. Courel, M., Clément, Y., Bossevain, C., Foretek, D., Vidal Cruchez, O., Yi, Z., *et al.* (2019) GC content shapes mRNA storage and decay in human cells. *elife* **8**, e49708
104. Wayment-Steele, H. K., Kim, D. S., Choe, C. A., Nicol, J. J., Wellington-Oguri, R., Watkins, A. M., *et al.* (2021) Theoretical basis for stabilizing messenger RNA through secondary structure design. *Nucleic Acids Res.* **49**, 10604–10617
105. Blenke, E. O., Örnskov, E., Schöneich, C., Nilsson, G. A., Volkin, D. B., Mastrobattista, E., *et al.* (2023) The storage and in-use stability of mRNA vaccines and therapeutics: not a cold case. *J. Pharm. Sci.* **112**, 386–403
106. Terai, G., Kamegai, S., and Asai, K. (2016) CDSfold: an algorithm for designing a protein-coding sequence with the most stable secondary structure. *Bioinformatics* **32**, 828–834

107. Zuker, M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science* **244**, 48–52
108. Zhang, H., Zhang, L., Lin, A., Xu, C., Li, Z., Liu, K., *et al.* (2023) Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature* **621**, 396–403
109. Zhang, H., Zhang, L., Liu, K., Li, S., Mathews, D. H., and Huang, L. (2022) Linear-time algorithms for RNA structure prediction. In: Kawaguchi, R. K., Iwakiri, J., eds. *RNA Structure Prediction*, Springer US, New York, NY: 15–34
110. Serra, M. J., and Turner, D. H. (1995) [11] Predicting thermodynamic properties of RNA. In *Methods in Enzymology*, 259th, Academic Press, Cambridge, MA: 242–261
111. Zuber, J., Schroeder, S. J., Sun, H., Turner, D. H., and Mathews, D. H. (2022) Nearest neighbor rules for RNA helix folding thermodynamics: improved end effects. *Nucleic Acids Res.* **50**, 5251–5262
112. Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., *et al.* (2011) ViennaRNA package 2.0. *Algorithms Mol. Biol.* **6**, 1–14
113. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415
114. Cao, S., and Chen, S. J. (2009) Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA* **15**, 696–706
115. Cheng, Y., Zhang, S., Xu, X., and Chen, S. J. (2021) Vfold2D-MC: a physics-based hybrid model for predicting RNA secondary structure folding. *J. Phys. Chem. B.* **125**, 10108–10118
116. Xu, X., Zhao, P., and Chen, S. J. (2014) Vfold: a web server for RNA structure and folding thermodynamics prediction. *PLoS One* **9**, e107504
117. Zhang, S., Li, J., and Chen, S. J. (2024) Machine learning in RNA structure prediction: advances and challenges. *Biophys. J.* **123**, 2647–2657
118. Muneer, A., Fati, S. M., Akbar, N. A., Agustriawan, D., and Wahyudi, S. T. (2022) iVaccine-Deep: prediction of COVID-19 mRNA vaccine degradation using deep learning. *J. King Saud Univ. Comput. Inf. Sci.* **34**, 7419–7432
119. Wayment-Steele, H. K., Kladwang, W., Watkins, A. M., Kim, D. S., Tunguz, B., Reade, W., *et al.* (2022) Deep learning models for predicting RNA degradation via dual crowdsourcing. *Nat. Mach. Intell.* **4**, 1174–1184
120. [preprint] He, S., Huang, R., Townley, J., Kretsch, R. C., Karagianes, T. G., Cox, D. B., *et al.* (2024) Ribonanza: deep learning of RNA structure through dual crowdsourcing. *bioRxiv*. https://doi.org/10.1101/2024.02.24.581671
121. Yit, T. W., Hassan, R., Zakaria, N. H., Kasim, S., Moi, S. H., Khairuddin, A. R., *et al.* (2023) Transformer in mRNA degradation prediction. *JOIV: Int. J. Inform. Visualization.* **7**, 588–599
122. Imran, S. A., Islam, M. T., Shahnaz, C., Islam, M. T., Imam, O. T., and Haque, M. (2020) COVID-19 mRNA vaccine degradation prediction using regularized LSTM model. In *2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, IEEE, New York City, NY: 328–331
123. Chze, O. N., and Abdullah, A. A. (2022) COVID-19 mRNA vaccine degradation prediction by using deep learning algorithms. In *2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, IEEE, New York City, NY: 444–450
124. Rabani, M., Levin, J. Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., *et al.* (2011) Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* **29**, 436–442
125. O'Brien, E. P., Ciryam, P., Vendruscolo, M., and Dobson, C. M. (2014) Understanding the influence of codon translation rates on cotranslational protein folding. *Acc. Chem. Res.* **47**, 1536–1544
126. Atkins, J. F., Loughran, G., Bhatt, P. R., Firth, A. E., and Baranov, P. V. (2016) Ribosomal frameshifting and transcriptional slippage: from genetic steganography and cryptography to adventitious use. *Nucleic Acids Res.* **44**, 7007–7078
127. Houseley, J., and Tollervey, D. (2009) The many pathways of RNA degradation. *Cell* **136**, 763–776
128. Mauger, D. M., Cabral, B. J., Presnyak, V., Su, S. V., Reid, D. W., Goodman, B., *et al.* (2019) mRNA structure regulates protein expression through changes in functional half-life. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 24075–24083
129. Verbeke, R., Hogan, M. J., Loré, K., and Pardi, N. (2022) Innate immune mechanisms of mRNA vaccines. *Immunity* **55**, 1993–2005
130. Deutscher, M. P., and Li, Z. (2001) Exoribonucleases and their multiple roles in RNA metabolism. *Prog. Nucleic Acid Res. Mol. Biol.* **66**, 67–105
131. Andrade, J. M., Pobre, V., Silva, I. J., Domingues, S., and Arraiano, C. M. (2009) The role of 3′–5′ exoribonucleases in RNA degradation. *Prog. Mol. Biol. Transl. Sci.* **85**, 187–229
132. Chang, J. H., Xiang, S., and Tong, L. (2011) 5′-3′ exoribonucleases. In: Nicholson, A. W., ed. *Ribonucleases*, Springer Berlin Heidelberg, Berlin, Heidelberg: 167–192
133. Li, W. M., Barnes, T., and Lee, C. H. (2010) Endoribonucleases–enzymes gaining spotlight in mRNA metabolism. *FEBS J.* **277**, 627–641
134. Li, Z., and Deutscher, M. P. (2004) Exoribonucleases and endoribonucleases. *EcoSal Plus* **1**, 10–1128
135. Keiler, K. C. (2015) Mechanisms of ribosome rescue in bacteria. *Nat. Rev. Microbiol.* **13**, 285–297
136. Collart, M. A., and Weiss, B. (2020) Ribosome pausing, a dangerous necessity for co-translational events. *Nucleic Acids Res.* **48**, 1043–1055
137. Yip, M. C., and Shao, S. (2021) Detecting and rescuing stalled ribosomes. *Trends Biochem. Sci.* **46**, 731–743
138. Juszkiewicz, S., Slodkowicz, G., Lin, Z., Freire-Pritchett, P., Peak-Chew, S. Y., and Hegde, R. S. (2020) Ribosome collisions trigger cis-acting feedback inhibition of translation initiation. *Elife* **9**, e60038
139. Saito, K., Kratzat, H., Campbell, A., Buschauer, R., Burroughs, A. M., Berninghausen, O., *et al.* (2022) Ribosome collisions induce mRNA cleavage and ribosome rescue in bacteria. *Nature* **603**, 503–508
140. Best, K., Ikeuchi, K., Kater, L., Best, D., Musial, J., Matsuo, Y., *et al.* (2023) Structural basis for clearing of ribosome collisions by the RQT complex. *Nat. Commun.* **14**, 921
141. Mitarai, N., Sneppen, K., and Pedersen, S. (2008) Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization. *J. Mol. Biol.* **382**, 236–245
142. Liutkute, M., Samatova, E., and Rodnina, M. V. (2020) Cotranslational folding of proteins on the ribosome. *Biomolecules* **10**, 97
143. Thommen, M., Holtkamp, W., and Rodnina, M. V. (2017) Co-translational protein folding: progress and methods. *Curr. Opin. Struct. Biol.* **42**, 83–89
144. Thanaraj, T. A., and Argos, P. (1996) Protein secondary structural types are differentially coded on messenger RNA. *Protein Sci.* **5**, 1973–1983
145. Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., *et al.* (2016) Synonymous codons direct cotranslational folding toward different protein conformations. *Mol. Cell.* **61**, 341–351
146. O'brien, E. P., Vendruscolo, M., and Dobson, C. M. (2012) Prediction of variable translation rate effects on cotranslational protein folding. *Nat. Commun.* **3**, 868
147. Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chételat, G., Teunissen, C. E., *et al.* (2021) Alzheimer's disease. *Lancet* **397**, 1577–1590
148. Yan, S., Zhu, Q., Jain, S., and Schlick, T. (2022) Length-dependent motions of SARS-CoV-2 frameshifting RNA pseudoknot and alternative conformations suggest avenues for frameshifting suppression. *Nat. Commun.* **13**, 4284
149. Cao, S., and Chen, S. J. (2008) Predicting ribosomal frameshifting efficiency. *Phys. Biol.* **5**, 016002
150. Boros, L. G., Kyriakopoulos, A. M., Brogna, C., Piscopo, M., McCullough, P. A., and Seneff, S. (2024) Long-lasting, biochemically modified mRNA, and its frameshifted recombinant spike proteins in human tissues and circulation after COVID-19 vaccination. *Pharmacol. Res. Perspect.* **12**, e1218
151. Mulroney, T. E., Pöyry, T., Yam-Puc, J. C., Rust, M., Harvey, R. F., Kalmar, L., *et al.* (2024) N 1-methylpseudouridylation of mRNA causes+1 ribosomal frameshifting. *Nature* **625**, 189–194
152. [preprint] Krueger, R., and Ward, M. (2024) Scalable differentiable folding for mRNA design. *bioRxiv*. https://doi.org/10.1101/2024.05.29.594436

153. Matthies, M. C., Krueger, R., Torda, A. E., and Ward, M. (2024) Differentiable partition function calculation for RNA. *Nucleic Acids Res.* **52**, e14
154. LeCun, Y., Bengio, Y., and Hinton, G. (2015) Deep learning. *Nature* **521**, 436–444
155. Tian, Y., Si, L., Zhang, X., Cheng, R., He, C., Tan, K. C., *et al.* (2021) Evolutionary large-scale multi-objective optimization: a survey. *ACM Comput. Surv.* **54**, 1–34
156. Chen, Y., Li, Y., Narayan, R., Subramanian, A., and Xie, X. (2016) Gene expression inference with deep learning. *Bioinformatics* **32**, 1832–1839
157. Ahmed, O., and Brifcani, A. (2019) Gene expression classification based on deep learning. In *2019 4th Scientific International Conference Najaf (SICN)*, IEEE, New York City, NY: 145–149
158. Xiao, Z., Li, W., Moon, H., Roell, G. W., Chen, Y., and Tang, Y. J. (2023) Generative artificial intelligence GPT-4 accelerates knowledge mining and machine learning for synthetic biology. *ACS Synth. Biol.* **12**, 2973–2982
159. Lopez, R., Gayoso, A., and Yosef, N. (2020) Enhancing scientific discoveries in molecular biology with deep generative models. *Mol. Syst. Biol.* **16**, e9198
160. Winnifrith, A., Outeiral, C., and Hie, B. L. (2024) Generative artificial intelligence for de novo protein design. *Curr. Opin. Struct. Biol.* **86**, 102794
161. Nachtergaele, S., and He, C. (2018) Chemical modifications in the life of an mRNA transcript. *Annu. Rev. Genet.* **52**, 349–372
162. Boo, S. H., and Kim, Y. K. (2020) The emerging role of RNA modifications in the regulation of mRNA stability. *Exp. Mol. Med.* **52**, 400–408
163. Nance, K. D., and Meier, J. L. (2021) Modifications in an emergency: the role of N1-methylpseudouridine in COVID-19 vaccines. *ACS Cent. Sci.* **7**, 748–756
164. Liu, A., and Wang, X. (2022) The pivotal role of chemical modifications in mRNA therapeutics. *Front. Cell Dev. Biol.* **10**, 901510
165. Gilbert, W. V., and Nachtergaele, S. (2023) mRNA regulation by RNA modifications. *Annu. Rev. Biochem.* **92**, 175–198
166. Cappannini, A., Ray, A., Purta, E., Mukherjee, S., Boccaletto, P., Moafinejad, S. N., *et al.* (2024) MODOMICS: a database of RNA modifications and related information. 2023 update. *Nucleic Acids Res.* **52**, D239–D244
167. Morais, P., Adachi, H., and Yu, Y. T. (2021) The critical contribution of pseudouridine to mRNA COVID-19 vaccines. *Front. Cell Dev. Biol.* **9**, 789427
168. Anderson, B. R., Muramatsu, H., Nallagatla, S. R., Bevilacqua, P. C., Sansing, L. H., Weissman, D., *et al.* (2010) Incorporation of pseudouridine into mRNA enhances translation by Diminishing PKR activation. *Nucleic Acids Res.* **38**, 5884–5892
169. Karikó, K., Muramatsu, H., Welsh, F. A., Ludwig, J., Kato, H., Akira, S., *et al.* (2008) Incorporation of pseudouridine into mRNA yields superior nonimmunogenic vector with increased translational capacity and biological stability. *Mol. Ther.* **16**, 1833–1840
170. Feldman, R. A., Fuhr, R., Smolenov, I., Ribeiro, A. M., Panther, L., Watson, M., *et al.* (2019) mRNA vaccines against H10N8 and H7N9 influenza viruses of pandemic potential are immunogenic and well tolerated in healthy adults in phase 1 randomized clinical trials. *Vaccine* **37**, 3326–3334
171. John, S., Yuzhakov, O., Woods, A., Deterling, J., Hassett, K., Shaw, C. A., *et al.* (2018) Multi-antigenic human cytomegalovirus mRNA vaccines that elicit potent humoral and cell-mediated immunity. *Vaccine* **36**, 1689–1699
172. Leal, L., Guardo, A. C., Morón-López, S., Salgado, M., Mothe, B., Heirman, C., *et al.* (2018) Phase I clinical trial of an intranodally administered mRNA-based therapeutic vaccine against HIV-1 infection. *Aids* **32**, 2533–2545
173. Meyer, M., Huang, E., Yuzhakov, O., Ramanathan, P., Ciaramella, G., and Bukreyev, A. (2018) Modified mRNA-based vaccines elicit robust immune responses and protect Guinea pigs from Ebola virus disease. *J. Infect. Dis.* **217**, 451–455
174. Pardi, N., Hogan, M. J., Pelc, R. S., Muramatsu, H., Andersen, H., DeMaso, C. R., *et al.* (2017) Zika virus protection by a single low-dose nucleoside-modified mRNA vaccination. *Nature* **543**, 248–251
175. Shaw, C., Lee, H., Knightly, C., Kalidindi, S., Zaks, T., Smolenov, I., *et al.* (2019) Phase 1 trial of an mRNA-based combination vaccine against hMPV and PIV3. *Open Forum Infect. Dis.* **6**, S970
176. Aditham, A., Shi, H., Guo, J., Zeng, H., Zhou, Y., Wade, S. D., *et al.* (2022) Chemically modified mocRNAs for highly efficient protein expression in mammalian cells. *ACS Chem. Biol.* **17**, 3352–3366
177. Chawla, M., Oliva, R., Bujnicki, J. M., and Cavallo, L. (2015) An atlas of RNA base pairs involving modified nucleobases with optimal geometries and accurate energies. *Nucleic Acids Res.* **43**, 6714–6729
178. McCown, P. J., Ruszkowska, A., Kunkler, C. N., Breger, K., Hulewicz, J. P., Wang, M. C., *et al.* (2020) Naturally occurring modified ribonucleosides. *Wiley Interdiscip. Rev. RNA* **11**, e1595
179. D'Esposito, R. J., Myers, C. A., Chen, A. A., and Vangaveti, S. (2022) Challenges with simulating modified RNA: insights into role and reciprocity of experimental and computational approaches. *Genes* **13**, 540
180. Varenyk, Y., Spicher, T., Hofacker, I. L., and Lorenz, R. (2023) Modified RNAs and predictions with the ViennaRNA package. *Bioinform* **39**, btad696
181. Hopfinger, M. C., Kirkpatrick, C. C., and Znosko, B. M. (2020) Predictions and analyses of RNA nearest neighbor parameters for modified nucleotides. *Nucleic Acids Res.* **48**, 8901–8913
182. Kierzek, E., Zhang, X., Watson, R. M., Kennedy, S. D., Szabat, M., Kierzek, R., *et al.* (2022) Secondary structure prediction for RNA sequences including N6-methyladenosine. *Nat. Commun.* **13**, 1271
183. Xu, X., Jin, L., Xie, L., and Chen, S. J. (2022) Landscape zooming toward the prediction of RNA cotranscriptional folding. *J. Chem. Theor. Comput.* **18**, 2002–2015
184. Ding, F., Sharma, S., Chalasani, P., Demidov, V. V., Broude, N. E., and Dokholyan, N. V. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA* **14**, 1164–1173
185. Boniecki, M. J., Lach, G., Dawson, W. K., Tomala, K., Lukasz, P., Soltysinski, T., *et al.* (2016) SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res.* **44**, e63
186. Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R., and Stadler, P. F. (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinf.* **9**, 1–13
187. Popenda, M., Szachniuk, M., Antczak, M., Purzycka, K. J., Lukasiak, P., Bartol, N., *et al.* (2012) Automated 3D structure composition for large RNAs. *Nucleic Acids Res.* **40**, e112
188. Rivas, E., and Eddy, S. R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* **285**, 2053–2068
189. Das, R., and Baker, D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 14664–14669
190. Parisien, M., and Major, F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452**, 51–55
191. Watkins, A. M., Rangan, R., and Das, R. (2020) FARFAR2: improved de novo rosetta prediction of complex global RNA folds. *Structure* **28**, 963–976
192. Andronescu, M. S., Pop, C., and Condon, A. E. (2010) Improved free energy parameters for RNA pseudoknotted secondary structure prediction. *RNA* **16**, 26–42
193. Ren, J., Rastegari, B., Condon, A., and Hoos, H. H. (2005) HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA* **11**, 1494–1504
194. Townshend, R. J., Eismann, S., Watkins, A. M., Rangan, R., Karelina, M., Das, R., *et al.* (2021) Geometric deep learning of RNA structure. *Science* **373**, 1047–1051
195. Do, C. B., Woods, D. A., and Batzoglou, S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22**, e90–e98

196. Zhang, D., Li, J., and Chen, S. J. (2021) IsRNA1: de novo prediction and blind screening of RNA 3D structures. *J. Chem. Theor. Comput.* **17**, 1842–1857
197. Zhang, D., Chen, S. J., and Zhou, R. (2021) Modeling noncanonical RNA base pairs by a coarse-grained IsRNA2 model. *J. Phys. Chem. B.* **125**, 11907–11915
198. Sato, K., Kato, Y., Hamada, M., Akutsu, T., and Asai, K. (2011) IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics* **27**, i85–i93
199. Li, J., and Chen, S. J. (2023) RNAJP: enhanced RNA 3D structure predictions with non-canonical interactions and global topology sampling. *Nucleic Acids Res.* **51**, 3341–3356
200. Huang, L., Zhang, H., Deng, D., Zhao, K., Liu, K., Hendrix, D. A., *et al.* (2019) LinearFold: linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search. *Bioinformatics* **35**, i295–i304
201. Li, J., Zhu, W., Wang, J., Li, W., Gong, S., Zhang, J., *et al.* (2018) RNA3DCNN: local and global quality assessments of RNA 3D structures using 3D deep convolutional neural networks. *PLoS Comput. Biol.* **14**, e1006514
202. Tan, Z., Fu, Y., Sharma, G., and Mathews, D. H. (2017) TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic Acids Res.* **45**, 11570–11581
203. [preprint] Zhang, S., Liu, Y., and Xie, L. (2022) Physics-aware graph neural network for accurate RNA 3D structure prediction. *arXiv. 10.48550/arXiv.2210.16392.* https://doi.org/10.48550/arXiv.2210.16392
204. Singh, J., Hanson, J., Paliwal, K., and Zhou, Y. (2019) RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat. Commun.* **10**, 5407
205. [preprint] Pearce, R., Omenn, G. S., and Zhang, Y. (2022) De novo RNA tertiary structure prediction at atomic resolution using geometric potentials from deep learning. *BioRxiv*. https://doi.org/10.1101/2022.05.15.491755
206. [preprint] Chen, X., Li, Y., Umarov, R., Gao, X., and Song, L. (2020) RNA secondary structure prediction by learning unrolled algorithms. *arXiv*. https://doi.org/10.48550/arXiv.2002.05810
207. Wang, W., Feng, C., Han, R., Wang, Z., Ye, L., Du, Z., *et al.* (2023) trRosettaRNA: automated prediction of RNA 3D structure with transformer network. *Nat. Commun.* **14**, 7266
208. Sato, K., Akiyama, M., and Sakakibara, Y. (2021) RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat. Commun.* **12**, 941
209. [preprint] Sha, C. M., Wang, J., and Dokholyan, N. V. (2022) Predicting 3D RNA structure from solely the nucleotide sequence using Euclidean distance neural networks. *bioRxiv*. https://doi.org/10.1101/2022.05.16.492153
210. Wayment-Steele, H. K., Kladwang, W., Strom, A. I., Lee, J., Treuille, A., Becka, A., *et al.* (2022) RNA secondary structure packages evaluated and improved by high-throughput experiments. *Nat. Methods* **19**, 1234–1242
211. [preprint] Shen, T., Hu, Z., Peng, Z., Chen, J., Xiong, P., Hong, L., *et al.* (2022) E2Efold-3D: end-to-end deep learning method for accurate de novo RNA 3D structure prediction. *arXiv*. https://doi.org/10.48550/arXiv.2207.01586
212. Fu, L., Cao, Y., Wu, J., Peng, Q., Nie, Q., and Xie, X. (2022) UFold: fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Res.* **50**, e14
213. [preprint] Baek, M., McHugh, R., Anishchenko, I., Baker, D., and DiMaio, F. (2022) Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA. *bioRxiv*. https://doi.org/10.1101/2022.09.09.507333
214. Wang, L., Liu, Y., Zhong, X., Liu, H., Lu, C., Li, C., *et al.* (2019) DMfold: a novel method to predict RNA secondary structure with pseudoknots based on deep learning and improved base pair maximization principle. *Front. Genet.* **10**, 143
215. Li, Y., Zhang, C., Feng, C., Pearce, R., Lydia Freddolino, P., and Zhang, Y. (2023) Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction. *Nat. Commun.* **14**, 5745
216. Saman Booy, M., Ilin, A., and Orponen, P. (2022) RNA secondary structure prediction with convolutional neural networks. *BMC Bioinf.* **23**, 58
217. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., *et al.* (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500
218. [preprint] Franke, J. K., Runge, F., and Hutter, F. (2023) Scalable deep learning for RNA secondary structure prediction. *arXiv*. https://doi.org/10.48550/arXiv.2307.10073
219. Zhang, Y., Wang, J., and Xiao, Y. (2022) 3dRNA: 3D structure prediction from linear to circular RNAs. *J. Mol. Biol.* **434**, 167452
220. Mao, K., Wang, J., and Xiao, Y. (2022) Length-dependent deep learning model for RNA secondary structure prediction. *Molecules* **27**, 1030
221. Li, J., Zhang, S., Zhang, D., and Chen, S. J. (2022) Vfold-Pipeline: a web server for RNA 3D structure prediction from sequences. *Bioinformatics* **38**, 4042–4043