scientific reports

OPEN

StructmRNA a BERT based model with dual level and conditional masking for mRNA representation

Sepideh Nahali[1,3]✉, Leila Safari[3], Alireza Khanteymoori[2] & Jimmy Huang[1]

In this study, we introduce StructmRNA, a new BERT-based model that was designed for the detailed analysis of mRNA sequences and structures. The success of DNABERT in understanding the intricate language of non-coding DNA with bidirectional encoder representations is extended to mRNA with StructmRNA. This new model uses a special dual-level masking technique that covers both sequence and structure, along with conditional masking. This enables StructmRNA to adeptly generate meaningful embeddings for mRNA sequences, even in the absence of explicit structural data, by capitalizing on the intricate sequence-structure correlations learned during extensive pre-training on vast datasets. Compared to well-known models like those in the Stanford OpenVaccine project, StructmRNA performs better in important tasks such as predicting RNA degradation. Thus, StructmRNA can inform better RNA-based treatments by predicting the secondary structures and biological functions of unseen mRNA sequences. The proficiency of this model is further confirmed by rigorous evaluations, revealing its unprecedented ability to generalize across various organisms and conditions, thereby marking a significant advance in the predictive analysis of mRNA for therapeutic design. With this work, we aim to set a new standard for mRNA analysis, contributing to the broader field of genomics and therapeutic development.

Keywords StructmRNA, mRNA sequence, Representation learning, mRNA structure, Dual-level masking, RNA degradation prediction, BERT

Bioinformatics, which combine machine learning with genomics, is driving breakthroughs. Using linguistic parallels in genetic sequences[1] and mRNA therapeutic advances against SARS-CoV-2[2], researchers have addressed key challenges, such as mRNA sequence representation learning and predicting sequence features. Predicting mRNA degradation is vital for understanding mRNA functionality, and crucial for biological processes such as vaccine development and therapeutic research. However, traditional statistical models and neural networks[3] struggle to capture semantic dependencies and long-range context in mRNA sequences. While there is growing interest in advanced neural networks, they face challenges such as data scarcity and limited sequence-structure understanding. Thus, the lack of high-quality mRNA datasets with detailed structural annotations hinders model progress.

To address these challenges, we introduce StructmRNA in this study. This computational model leverages the Bidirectional Encoder Representations from Transformers (BERT) framework[4,5], traditionally a linchpin of natural language processing (NLP), to interpret and predict the nuanced language of mRNA sequences and structures. BERT's bidirectional context analysis is excellent for capturing nuanced language contexts, which is analogous to understanding nucleotide dynamics in mRNA. Thus, StructmRNA predicts mRNA structures and functions robustly, even without explicit structural data. StructmRNA, as a representation learning method, utilizes BERT for mRNA sequences and structures. It employs dual-level masking to enhance mRNA representation. Inspired by BERT's training, it deciphers complex mRNA relationships, advancing bioinformatics and therapeutic research. However, the novelty of StructmRNA lies in its integration of mRNA sequence and structural data for representation learning, using a dual-level masking technique and trained on a large dataset of sequences and structures. This approach enables improved accuracy and versatility in RNA-related tasks, especially in RNA degradation prediction. Its precision expedites the development of mRNA-based treatments and vaccines development, which is crucial for enabling rapid responses to emerging infectious diseases[6,7].

The impact of StructmRNA goes beyond therapeutics, significantly advancing mRNA virus research by elucidating mRNA structure-function relationships and aiding in synthetic mRNA design[8]. This versatility,

[1]Information Retrieval and Knowledge Management Research Lab, York University, Toronto, Ontario, Canada. [2]Department of Psychology, University of Freiburg, Freiburg, Germany. [3]Department of Computer Engineering, University of Zanjan, Zanjan, Iran. ✉email: sepidnah@yorku.ca

particularly in personalized medicine[9] and cross-species generalization, positions StructmRNA as a key player in shaping the future of molecular biology and medicine. Furthermore, integrating StructmRNA with extracellular vesicle (EV) RNA data could greatly enhance RNA research. StructmRNA's capabilities in mRNA structure-sequence representation learning offer promising applications in conjunction with resources like exoRBase 2.0, which supports the study of long RNAs (exLRs) from biofluids and advances biomarker discovery[10–12]. Characterizing EV RNA species has implications for disease mechanisms[13], with exLR signatures improving cancer diagnostics, including in SCLC and CRC[14,15]. Profiling cancer subtypes via EV-derived RNAs, such as in PDAC, can refine prognostics[16]. EV-origin analysis reveals patterns that reflect disease progression[17], supporting EV integration into StructmRNA for diagnostics. Further research is needed to enhance StructmRNA with EV data.

In this study, we conducted a case study on mRNA degradation prediction, demonstrating StructmRNA's adaptability. This highlights the potential of interdisciplinary approaches to uncover the complexities of biological systems.

Literature review

mRNA molecules play a crucial role in various biological processes, including gene expression, regulation, and viral pathogenesis. Traditional However, complete structural data are often needed for traditional mRNA analysis model's[18], which limits their application in analyzing novel mRNA sequences; additionally they often lack the predictive power of modern machine learning models. Thus, innovative computational methods have become essential for overcoming these challenges[19]. Specifically, the development of a model that can infer mRNA structure complexity and function with the mRNA sequence in the absence of explicit structural data is critical.

The evolution of computational RNA analysis methods has been well documented in recent works[20]. These methods, while innovative, often struggle with the complexity and variability of mRNA structures. Recent advancements in machine learning, particularly deep learning, have opened new avenues for addressing these challenges, providing a significant shift from traditional sequence analysis methods[4,21]. The field of representation learning has grown significantly over the past decade[22]. Notable advancements have been made in sequence analysis techniques, particularly for those for RNA[23]. The use of sequence-to-sequence autoencoder models[24], CNNs, LSTMs[25], Variational Autoencoders (VAEs)[26], and Graph Neural Networks (GNNs)[27,28] demonstrates the diversification of approaches in understanding the complexity of RNA structures. Additionally, transformer-based models, such as BERT, have revolutionized the analysis of biological sequences, including DNA and proteins, although they do not account for structural information[1,26]. Moreover, techniques such as dna2vec and rna2vec have been specifically developed for gene embeddings[29]. Embedding methods, including adaptations of NLP techniques such as Word2Vec, have significantly advanced sequence representation in bioinformatics[30–32].

CNNs are excellent at detecting local sequence patterns, while attention graph convolutions (AGCs) harness graph structures to capture intricate RNA relationships. Transformers, such as BERT excel at sequence tasks due to self-attention, which captures long-range dependencies better than LSTMs can[33,34], making BERT ideal for mRNA sequence analysis. However, many models overlook mRNA secondary structure, reducing performance in RNA degradation prediction[35]. The advancements in bioinformatics have highlighted the potential of the abovementioned approaches in predicting mRNA degradation, a crucial factor in designing stabilized RNA therapeutics[36–38]. RNA degradation prediction is crucial for mRNA stability, therapeutic applications, gene expression, and viral RNA research and is impacted by RNA secondary structure, with specific motifs affecting rates with recent advances, sequence and structural data have been integrated to improve prediction accuracy, as in COVID-19 vaccine mRNA stability models.

However, scalable, generalizable models that combine both sequence and structure are still lacking. Limited data, especially for specific mRNA types, challenges model generalizability, and public datasets such as Stanford's OpenVaccine are often small and diverse. The OpenVaccine Kaggle competition[39], a collaborative effort involving 1636 teams, aimed at predicting RNA degradation rates using computational models. This competition highlighted a variety of approaches, including linear regression models such as DegScore[40] and a version of the DegScore featurization with XGBoost[41], graph-based distance embeddings, and complex architectures combining autoencoders, GNNs, GRUs, and CNNs[2].

Traditional degradation prediction methods, such as those based on one-hot encoding, work for small-scale problems but struggle with biological relevance, high dimensionality, and scalability, so they are unsuitable for complex RNA sequences[42]. Techniques such as DegScore use statistical models with handcrafted mRNA features, which oversimplify nucleotide-degradation relationships and fail to capture RNA secondary structures vital for stability. These methods are limited in flexibility and scalability, especially for large, diverse datasets in RNA therapeutics and viral RNA research. Models such as CNNs and GCNs, including Nullrecurrent from Kaggle's OpenVaccine[2], excel with local features but struggle with global context. Kazuki2, using LSTMs, GRUs, and CNNs, captures both local and long-range dependencies but faces high computational costs and vanishing gradients. These models predict RNA degradation but struggle with dataset-specific structures, so generalizability is limited. Transformer models such as BERT could enhance RNA degradation prediction using representation learning that captures sequence, structure, and their dependencies, leading to more refined results.

Data augmentation techniques such as noise injection offer limited gains due to simple sequences[43,44]. Generative Adversarial Networks (GANs), which generate more realistic synthetic sequences[45], show promise but are underexplored in relation to mRNA. Though effective in protein and virus generation, concerns remain about the biological viability and functionality of synthetic mRNA sequences.

An in-depth analysis of existing works reveals the following key challenges in mRNA representation learning and analysis: 1. Complexity and Variability: the complexity of mRNA sequences hinders traditional models (e.g., CNNs, RNNs) from capturing crucial long-range contextual information[20,35]. 2. Data Limitations: Existing methods rely on existence of structural data, limiting their applicability to novel mRNA sequences[18]. Thus, a model that can infer structure and function without explicit structural data is needed[19]. 3. Integration of Machine Learning: The effective integration of machine learning with genomics is essential yet developing, so innovative approaches are needed. 4. Advanced Computational Tools: More sophisticated computational tools are needed for the design and comprehension of evolving mRNA therapeutics[2,26]. 5. Interdisciplinary Challenges: Leveraging insights across domains, such as applying NLP techniques in bioinformatics, presents challenges[1,29]. The advanced training techniques employed in StructmRNA are a response to the challenges. By implementing a dual-level masking process, StructmRNA addresses the need for a more nuanced understanding of mRNA sequence-structure relationships. StructmRNA's ability to represent mRNA sequence and structure enhances the predictive power of downstream tasks, such as mRNA degradation prediction without explicit structural data. This has profound implications in various fields, including therapeutic research, vaccine development, and the study of mRNA viruses.

Methodology

StructmRNA uses advanced computational techniques to analyze and embed mRNA sequences and structures. It employs a comparative framework for RNA degradation prediction using BERT with a dual-level masking strategy. This approach includes masking thresholds, model architecture, training protocols, dataset configuration, and data loader setup to enhance mRNA sequence analysis.

Dual-level masking process

The dual-level masking process in StructmRNA integrates sequence and structural data for accurate mRNA sequence embeddings. This section details sequence-level masking, and structure-level masking. Sequence-level masking is inspired by BERT, in which nucleotides are randomly replaced by a masking token, prompting the model to predict them based on the surrounding context and learn sequence dependencies. It is grounded in the methodology of[4]. Complementing sequence-level masking, structure-level masking targets elements of the mRNA structure. This approach helps the model learn how sequences fold into structural motifs, highlighting the role of structural context in understanding mRNA function. We set a 25% masking probability for each nucleotide or structural element to balance uncertainty with informative data. Our random sequence masking strategy evaluates each nucleotide against a random number for masking, which can be described as follows:

For a sequence of nucleotides $S = \{s_1, s_2, \ldots, s_n\}$, each nucleotide s_i is compared against a randomly generated number r_i uniformly distributed between 0 and 1. If $r_i < p$ (where *p* is the masking probability), then s_i is replaced with a [MASK] token. This is formalized as Eq. (1):

$$s'_i = \begin{cases} \texttt{[MASK]} & \text{if } r_i < p, \\ s_i & \text{otherwise.} \end{cases} \tag{1}$$

Advanced masking techniques like conditional and dynamic pattern masking address nucleotide-specific significance and replicate RNA variability. There is a moderate positive correlation between sequence and structure masking, indicating that increased sequence masking often leads to increased structure masking, emphasizing the need to integrate both in modeling. The Pearson correlation coefficient $\rho_{\text{seq, struct}}$ is computed as Equation (2).

$$\rho_{\text{seq, struct}} = \frac{\text{cov}(\text{seq}_D, \text{struct}_D)}{\sigma_{\text{seq}_D} \cdot \sigma_{\text{struct}_D}}, \tag{2}$$

where cov represents covariance, and σ denotes standard deviation. The model's interdependence shows its ability to predict masked parts using context, improving generalization. With dual-level masking and mRNA-specific complexities, this model identifies key patterns such as secondary structure motifs, regulatory elements, splice sites, codon biases, and degradation signals. This capability facilitates RNA structure prediction from sequences alone, which is crucial when structural data is missing, and enhances mRNA sequence and structure analysis to provide a better understanding of their functional roles.

Figure 1 illustrates the dual-level masking process applied to a sample mRNA sequence, showcasing the approach we employ to mimic the natural variability in RNA sequences.

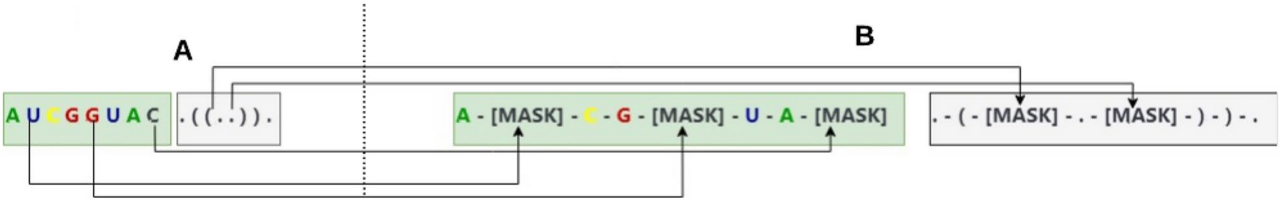

A
B
A U C G G U A C
. ((. .)) .
A - [MASK] - C - G - [MASK] - U - A - [MASK]
. - (- [MASK] - . - [MASK] -) -) - .

Fig. 1. A sample mRNA sequence and structure after dual-level masking. (**A**) Before masking. (**B**) After masking.

Conditional masking

This technique employs variable masking likelihood based on nucleotide type, facilitating conditional masking tailored to molecular structures and functions. It selectively targets nucleotides such as guanine, which are crucial for stability and function, reflecting their biological significance and variability in RNA sequences. This approach enhances realism by simulating natural variability observed in RNA sequences. To formalize this process, we introduce a function $P(s_i)$ that defines the probability of masking for each nucleotide type. For example, for guanine (G), we have the following relationship between the probabilities: $P(G) > P(A) = P(C) = P(U)$. The masking decision for a nucleotide s_i is then based on whether a random number r_i is less than $P(s_i)$.

Data preparation and processing pipeline for StructmRNA

In applying dual-level masking to our RNA dataset, we generate two key columns: masked_sequence and masked_structure, containing modified RNA sequences and structures. Both use the same masking token. We use the BERT tokenizer to map RNA sequences into token formats for training and prediction. Additionally, we developed a custom PyTorch Dataset class, RNADataset, to manage "PyTorch mRNA Dataset," which was specifically designed for our mRNA data. It handles masked sequences and structures to align seamlessly with BERT model requirements. To optimize training, we integrate a DataLoader with a custom collate function for batch-wise processing of tokenized RNA sequences and structures, ensuring efficient grouping while preserving BERT input integrity. We use a batch size of 16 to balance computational efficiency with learning capability. Larger batches might speed up training but reduce learning detail, while smaller ones slow training. This data configuration supports streamlined, effective training, ensuring accurate and efficient model predictions.

Tokenizer Configuration In our study, we developed a tokenization method for RNA sequence and structural data. Each nucleotide and structural symbol is converted into unique numerical identifiers using a custom dictionary, token2int, which includes a special [MASK] token. This [MASK] token is crucial for training, akin to BERT's masked language modeling, enabling context-based prediction. This method bridges RNA sequence complexity with transformer models, ensuring effective model training.

We optimized the hyperparameters of the StructmRNA model to improve prediction accuracy, as measured by MCRMSE, while ensuring efficient training. We performed automatic hyperparameter tuning using a grid search and conducted an ablation study to evaluate the importance of various model components. The optimal settings were as follows: hidden layer size 256, 8 layers, 8 attention heads, and intermediate layer size 500. More layers or attention heads offered minimal MCRMSE improvement but increased training time. A vocabulary size over 800 led to overfitting, longer training times, and higher memory use. The AdamW initial learning rate of 1e-5, OneCycleLR max learning rate of 1e-4, and 50 epochs with early stopping yielded the best results. This tuning ensures optimal performance and efficiency. The specific hyperparameters are highlighted in Table 1.

Figure 2 presents the flowchart of the data configuration and model training process used in our StructmRNA research. It begins with the "Original RNA dataset," which undergoes a "Masking Process" to generate the mentioned data columns, masked_sequence and masked_structure. These modified columns simulate scenarios in which certain nucleotides or structural elements are unknown, thus providing a realistic training environment for our model. "BERT Tokenization" follows this masking process and breaks down sequences and structures into forms that are usable for model training and prediction. The tokenized data are then managed within a custom PyTorch Dataset Class that has been specifically designed to handle the complexities of RNA data and facilitate efficient management during the training phase. The DataLoader, set with a batch size of 16, processes the "Pytorch mRNA Dataset" in batches using a custom collate function, optimizing the batch-wise processing and maintaining the integrity of the sequences. The final step, "Model Training," involves training the BERT-based deep learning model using the prepared data, translating computational preparations into practical outcomes and advancing our understanding of RNA degradation mechanisms. Figure 3 illustrates the architecture of the BERT model used in the StructmRNA, designed for sequence and structural prediction in a masked language modeling context. It shows the progression from the input of original sequences, through embedding layers and multiple transformer blocks to the final prediction of masked tokens.

Data augmentation with generative adversarial networks

In bioinformatics, limited datasets constrain predictive models. StructmRNA leverages GANs to augment data by replicating real mRNA sequences' statistical properties, enriching datasets with diverse samples. In this

Model	Hyperparameters
StructmRNA	Vocabulary size: 800, Hidden layer size: 256, Hidden layers count: 8, Attention heads: 8, Intermediate layer size: 500, AdamW initial learning rate: 1e-5, OneCycleLR max learning rate: 1e-4, Training epochs: 50, optimizer: AdamW, Loss function: CrossEntropyLoss, Early stopping patience: 5
Word2Vec	Window Size: 5, Embedding Dimensions: 50
ELMo	Number of Layers: 3, Hidden Units: 256, Activation Function: tanh, Dropout Rate: 0.4, Transformation Layer: sigmoid
LSTM	Number of Hidden Units: 256, Activation Function: Swish, Number of Layers: 3, Dropout Rate: 0.4
CNN	Kernel Size: 3, Filters per Layer: mean, Activation Function: ReLU, Pooling Size: 300
VAE	Latent Dimension: 256, 128, 64, 32, 16, Activation Function: LeakyReLU and sigmoid, Dropout Rate: 0.3, 0.2, 0.1
AGC	Number of Filters: 256, Filter Size: 7, Number of Layers: 4, Dropout Rate (embedding layer): 0.6/0.4

Table 1. Hyperparameters and training parameters for the StructmRNA model and various baseline models utilizing embedding methods for the RNA degradation prediction task over 400 training epochs.

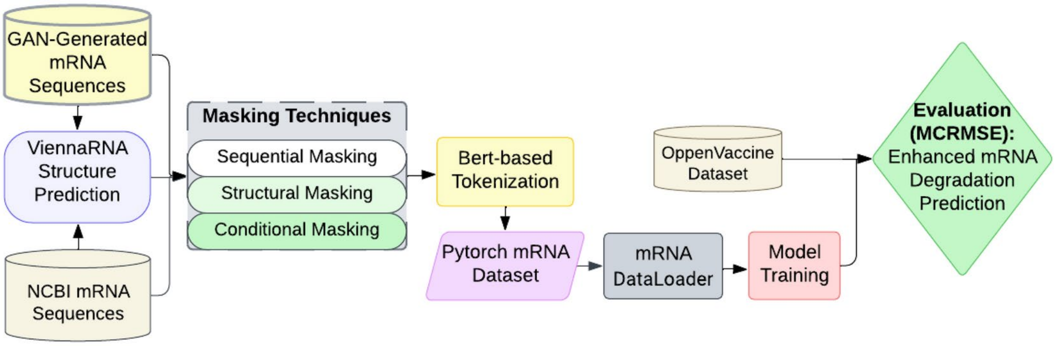


Fig. 2. StructmRNA pipeline from mRNA generation to model evaluation. NCBI and GAN-generated sequences undergo structure prediction via ViennaRNA, followed by sequential, structural, and conditional masking. Tokenized data is organized into a PyTorch dataset, processed through a DataLoader, and used for model training. Evaluation uses the OpenVaccine dataset with MCRMSE for mRNA degradation prediction.

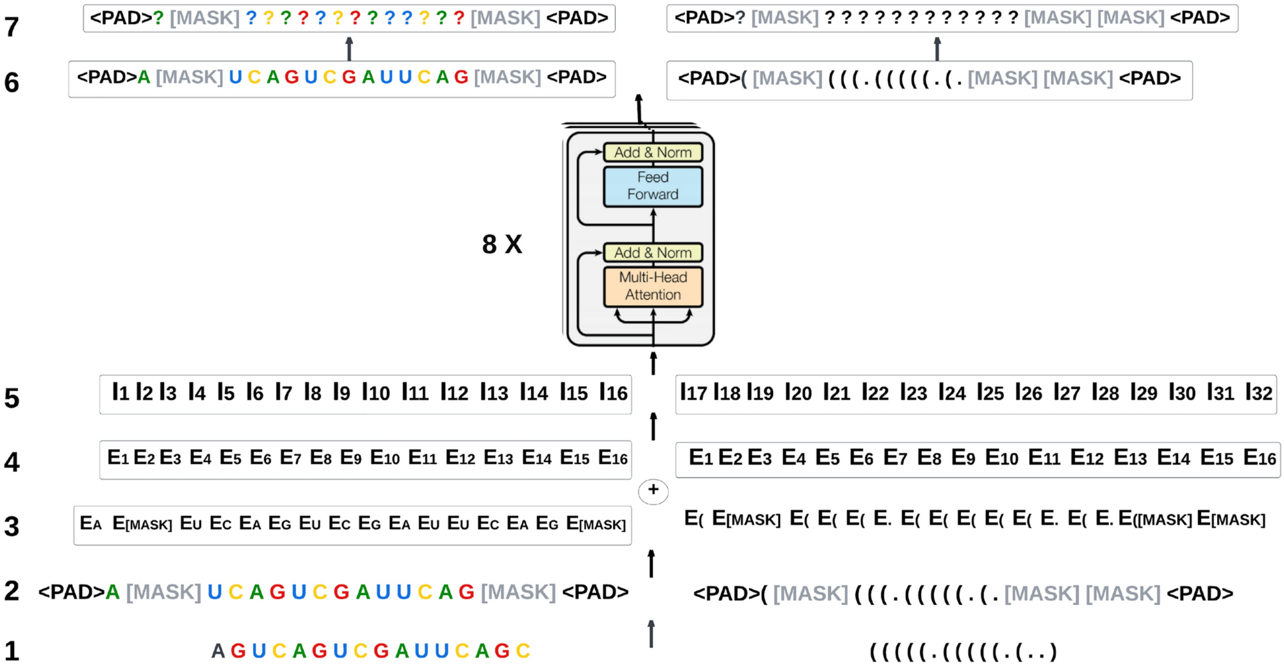

7
<PAD>? [MASK] ? ? ? ? ? ? ? ? ? ? ? ? ? ? [MASK] <PAD>
<PAD>? [MASK] ? ? ? ? ? ? ? ? ? ? ? ? [MASK] [MASK] <PAD>
6
<PAD>A [MASK] U C A G U C G A U U C A G [MASK] <PAD>
<PAD>([MASK] (((. (((((. (. [MASK] [MASK] <PAD>
8 X
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
5
I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 I13 I14 I15 I16
I17 I18 I19 I20 I21 I22 I23 I24 I25 I26 I27 I28 I29 I30 I31 I32
4
E1 E2 E3 E4 E5 E6 E7 E8 E9 E10 E11 E12 E13 E14 E15 E16
E1 E2 E3 E4 E5 E6 E7 E8 E9 E10 E11 E12 E13 E14 E15 E16
+
3
EA E[MASK] EU EC EA EG EU EC EG EA EU EU EC EA EG E[MASK]
E(E[MASK] E(E(E(E. E(E(E(E(E(E. E(E. E([MASK] E[MASK]
2
<PAD>A [MASK] U C A G U C G A U U C A G [MASK] <PAD>
<PAD>([MASK] (((. (((((. (. [MASK] [MASK] <PAD>
1
A G U C A G U C G A U U C A G C
(((((. (((((. (. .)

Fig. 3. StructmRNA's sequence and structure masking process: (1) Original sequence and structure, (2) Masked, (3) Token embedding, (4) Positional embedding, (5) Concatenation, (6) MLM prediction, (7) Predicted vs. original tokens.

way, dataset scarcity is addressed, and synthetic sequences in research are explored. GAN-generated sequences enhance training data volume and model generalization, improving robustness and biological relevance. Combining BERT's context-sensitive learning with GANs' data augmentation makes StructmRNA a pioneering advancement in bioinformatics, highlighting the potential of interdisciplinary strategies in analyzing mRNA sequences and structures. The use of GANs in StructmRNA raises concerns about the biological viability of generated sequences. Rigorous validation is crucial for ensuring these sequences are statistically accurate and biologically plausible[46]. Examining the biological significance of GAN-generated sequences highlights our commitment to responsibly and effectively harnessing the full potential of synthetic biology[45,47].

We chose a transformer-based GAN because it can handle sequential data with self-attention, which is crucial for mRNA sequences. It maintains nucleotide order and sequence structure through positional encoding, enhancing biological plausibility over simpler GANs such as CycleGAN. Figure 4a shows the process of integrating GAN data augmentation into our StructmRNA model. Figure 4b details the generator and discriminator architecture in the transformer GAN framework for synthetic mRNA sequence generation.

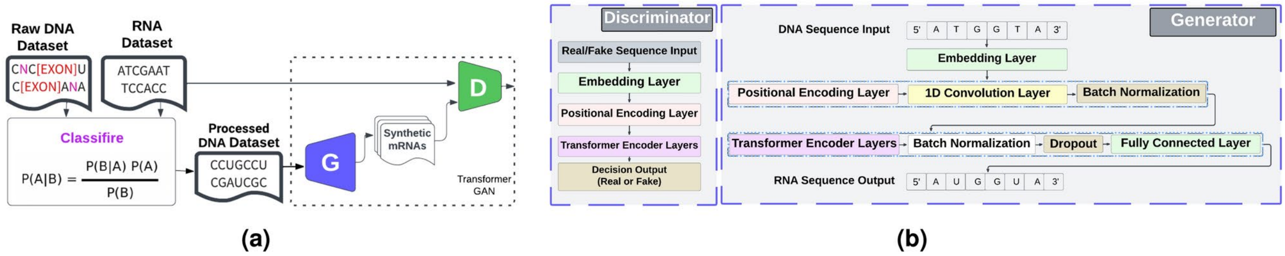


Fig. 4. (**a**) Workflow diagram for augmenting mRNA sequence and structure: Train mRNA classifier, apply to training set, generate synthetic sequences with transformer GAN, evaluate with classifier. (**b**) Generator and discriminator architecture of the transformer GAN for synthetic mRNA generation.

Experiments

Our study provides a comprehensive exploration of computational strategies for RNA degradation prediction, spanning from traditional statistical model DegScore[2] to more sophisticated neural network architectures such as CNNs, LSTMs, and transformer-based models (Table 1). We benchmark the StructmRNA model against various baseline models with different embedding and modeling techniques. This comparison validates our findings and highlights the advantages and limitations of each method for RNA degradation prediction. Our goal is to establish StructmRNA as a significant advancement in bioinformatics, offering improved predictive capabilities for mRNA-related tasks. In this study, we explored RNA degradation prediction using various computational strategies: DegScore, DegScore-XGBoost, Nullrecurrent, Kazuki2, Genetic algorithm, and Ensemble top two models[2]. Moreover we compared StructmRNA model with Word2vec, ELMo, LSTM, CNN, VAE, and AGC models (Table 1) to highlight its strengths and improvements. To validate StructmRNA, we used cross-validation with 80% of the data for training and 20% for validation. Performance was assessed using MCRMSE and ANOVA tests to analyze significant differences in loss among models. This metric ensures direct comparison with established benchmarks. These methods confirmed the robustness and comparative of StructmRNA in predicting advantages in RNA degradation.

Before passing the data to StructmRNA, we handled missing or incomplete mRNA structural information by applying a linear regression model to predict and impute the gaps. To achieve this, RNA sequences and structures were transformed into numerical formats. This imputation technique is robust for real-world datasets and improves the performance of predictive models like StructmRNA when dealing with incomplete annotations.

Generative adversarial network design

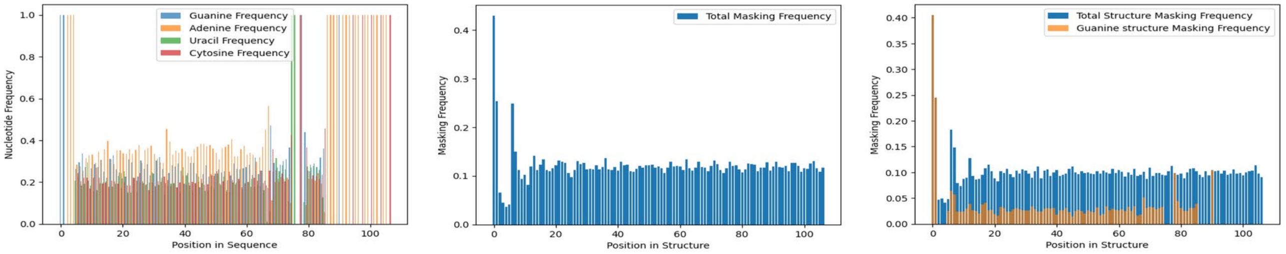
For our GAN architecture, we used PyTorch and included a Generator with 256-dimensional embeddings, positional encoding, and a transformer encoder with 10 layers and 64 heads (automatic parameter tuning with grid search). The Discriminator evaluated synthetic mRNA sequences versus real ones. A Naive Bayes classifier, trained with k-mer size 6, distinguished mRNA-like from non-coding sequences. mRNA-like sequences were concatenated, non-coding DNA was converted to mRNA-like sequences, and both were combined with real mRNA to enrich the training dataset.

Dataset

We built the StructmRNA model with 3.46 billion nucleotides of human mRNA data from NCBI and added 2,400 sequences from the Stanford OpenVaccine project to improve accuracy, especially for mRNA viruses. Vienna mRNA package annotations were used to enhance training and evaluation.

The BERT training dataset features nucleotide sequences from NCBI with characters 'A', 'U', 'G', and 'C', and secondary structure annotations using '(', ')', and '.'. For the mRNA degradation prediction task, each sequence in the dataset is accompanied by several key attributes, including a unique identifier (id), nucleotide sequence (sequence). Additionally, numerical arrays such as reactivity, deg_pH10, deg_Mg_pH10, deg_50C, and deg_Mg_50C represent the degradation likelihood under different experimental conditions, which is essential for understanding RNA degradation behavior. Error values associated with these measurements are provided under the error attribute. For more details on the data format, please refer to the Supplementary Materials. To ensure processing efficiency and improve model training, we standardized all sequences to 107 nucleotides. This uniform length facilitated streamlined training. Additionally, a GAN model was used to generate mRNA sequences, and synthetic sequences from NCBI DNA sources were added to enhance dataset diversity and robustness. We utilized 1.125 million nucleotides from human DNA, sourced from NCBI[48] and Stanford COVID Vaccine dataset[39], categorized into non-coding and coding segments. Then, we segmented mRNA sequences, including those from NCBI, into 107-nucleotide chunks using standard bioinformatics methods. To ensure data reliability, we applied stringent quality controls, including error filtering and sequence validation, especially on degradation-related attributes. Sequences that did not meet specific thresholds for quality and noise were excluded from the final dataset. This ensured that only high-confidence sequences were used to train the model. Additionally, the structural annotations were generated using established bioinformatics tools such as the ViennaRNA package, further enhancing the reliability of the dataset.

Figure 5 provides a comprehensive visual overview of the data preparation process that was central to our study on RNA sequences and structures. The first subfigure (Fig. 5a) presents the distribution of nucleotides



(a) Nucleotide frequency across sequences.

(b) Distribution of masking across sequences.

(c) Distribution of masking across RNA structure

Fig. 5. Comprehensive overview of the data preparation process for RNA sequences and structures.

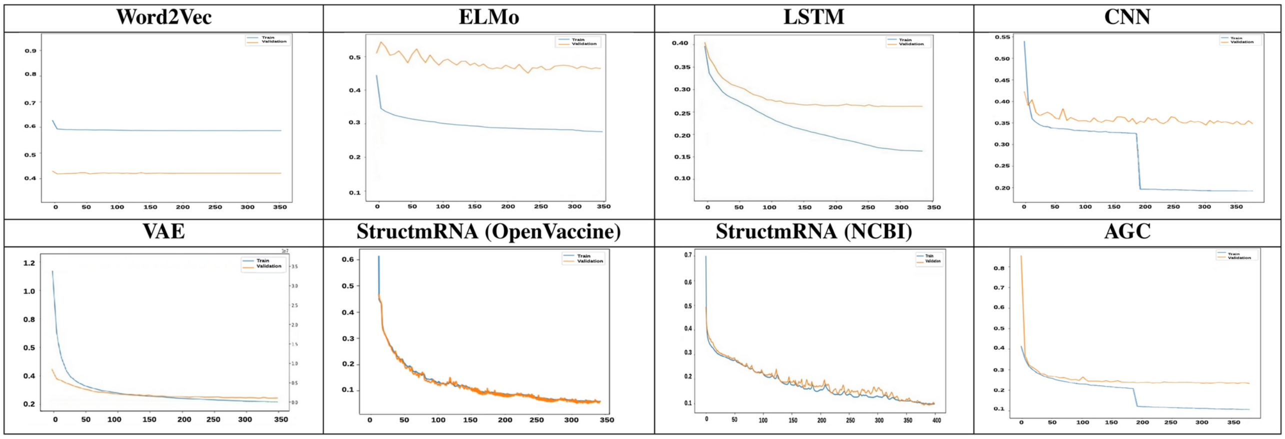


Fig. 6. StructmRNA performance in RNA degradation prediction. The figure shows training and validation MCRMSE losses (Y-axis) over 400 epochs (X-axis) for five target metrics averaged across four folds.

across different sequences, highlighting the variability and common patterns that our model must learn to recognize. The second (Fig. 5b) and third (Fig. 5c) subfigures depict the frequency distribution of masking in RNA sequences and structure positions, respectively, providing insights into how masking varies across different structural components. These visuals not only provide a better understanding of the complexity of the data but also highlight the rigor of our methodological framework for RNA degradation prediction (Fig. 6).

Results

We rigorously tested our StructmRNA model against baseline models, include Ensemble models, Genetic algorithm, Nullrecurrent, Kazuki2, DegScore-XGBoost, and DegScore from kaggle openVaccine competition, using the OpenVaccine dataset[2]. We also added some baseline methods focusing on their mRNA embedding skills with various neural networks (e.g., Word2Vec, ELMo, CNN, LSTM, VAE, and AGC), and top-performing Kaggle entries served as our reference points. We assessed our model across diverse baselines, using MCRMSE to measure RNA degradation prediction accuracy. This metric ensures direct comparison with established benchmarks. Table 2 summarizes the performance of these models on the OpenVaccine dataset, including results from our StructmRNA model. In our study, we evaluated the structmRNA model against various machine learning models based on three criteria: (1) improvement in training and Public Test set losses, (2) absolute difference (generalization gap) between end training and Public Test set losses, and (3) lowest achieved end losses. Findings from the OpenVaccine dataset highlight advancements in RNA degradation prediction across various modeling techniques.

Significantly, the **StructmRNA + OpenVaccine_Data** model, which utilizes the StructmRNA architecture pre-trained on OpenVaccine mRNA sequences along with a secondary structure dataset, demonstrates superior performance compared to alternative models with the lowest MCRMSE of **0.07**, indicating superior predictive accuracy. Following closely behind is **StructmRNA + GAN_Data**, along with **StructmRNA + NCBI_Data**, demonstrating the efficacy of integrating embedding methods with StructmRNA. These models achieved impressive MCRMSE scores of **0.11** and **0.10**, respectively. **StructmRNA + GAN_Data** utilizes the StructmRNA framework, which is pre-trained on synthetic mRNA sequences generated by GAN, combined with secondary structures provided by the Vienna tool. On the other hand, **StructmRNA + NCBI_Data** leverages the StructmRNA architecture, pre-trained on mRNA sequences sourced from NCBI, in conjunction with a secondary structure dataset. Among the models from the Kaggle OpenVaccine Competition, the Genetic

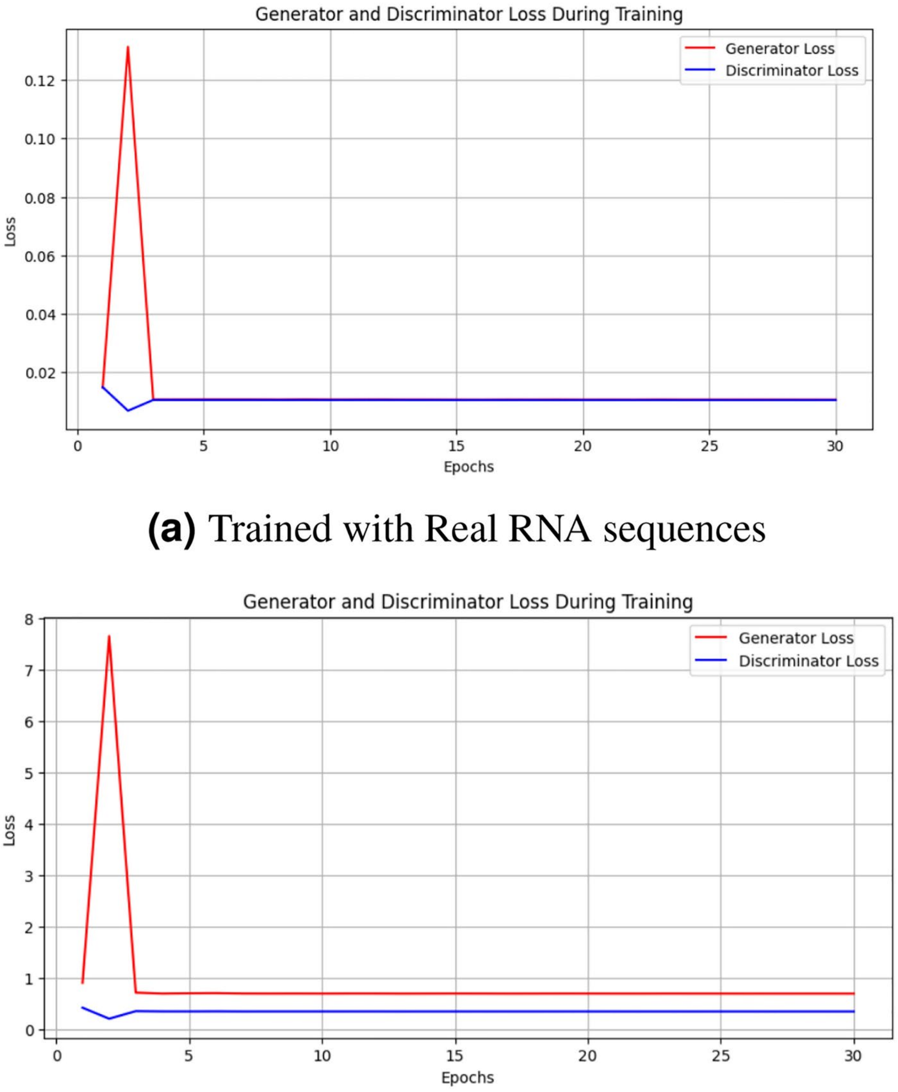
Models	**MCRMSE (Public test set)**
Models From Kaggle OpenVaccine Competition	
Experimental error	0.12
DegScore	0.39
DegScore-XGBoost	0.36
Nullrecurrent	0.23
Kazuki2	0.23
Genetic algorithm (10 of top 100 selected)	0.22
Ensemble top two models	0.22
Models in respect to embedding methods	
Word2vec	0.41
ELMo	0.44
LSTM	0.26
CNN	0.35
VAE	0.21
AGC	0.25
StructmRNA + OpenVaccine_Data	**0.07**
StructmRNA + NCBI_Data	**0.10**
StructmRNA + GAN_Data	**0.11**

Table 2. Performance of RNA degradation models on OpenVaccine dataset, including StructmRNA pretrained on OpenVaccine, NCBI, and GAN datasets (rounded to two decimals). Significant values are in bold.

Algorithm and Kazuki2 perform competitively, with MCRMSE scores of 0.22 and 0.23, respectively. The Genetic Algorithm, focused on optimization rather than deep learning, shows moderate accuracy but struggles with generalizing to complex RNA structures. Kazuki2, leveraging LSTMs, CNNs, and GRUs, captures both local and global dependencies. However, its higher computational costs and the risk of vanishing gradients contribute to its slightly elevated MCRMSE. The combination of CNNs and GCNs enables Nullrecurrent to achieve an MCRMSE of 0.23, indicating better accuracy compared to CNN-only models. The GCN's ability to model the secondary structure adds an extra layer of context that improves predictions, especially in tasks such as RNA degradation, where both sequence and structure play a vital role. Traditional embedding models such as Word2Vec and ELMo register higher MCRMSE values (0.41 and 0.44). While these methods are effective for use in NLP, they cannot account for the critical secondary structure and sequence of RNA molecules, and they require more data to perform well. This reflects their limitations in RNA degradation prediction. Similarly, DegScore and DegScore-XGBoost, though effective with small-scale features, oversimplify RNA sequence, structure, and nucleotide relationships, and rely on fixed feature sets, resulting in only moderate performance. Other neural networks such as CNNs, LSTMs, and VAEs improve on these methods, with MCRMSE values ranging from 0.35 to 0.26 and 0.21. However, they still fall short due to limited structural integration and also require more data to perform well. AGC performs better by capturing RNA secondary structures but faces computational challenges. Figure 6 shows training and validation losses of various sequence embedding models for predicting RNA degradation. It evaluates five metrics-reactivity, deg Mg pH10, deg pH10, deg Mg 50C, and deg 50C-averaged across four folds, comparing predicted and actual values.

Furthermore, although pre-training the StructmRNA with synthetic data does not improve the results of the RNA degradation task, the convergence observed in our customized BERT model (Fig. 7) shows that both models achieve similar convergence patterns within 30 epochs, indicating the model's robustness and the synthetic data's fidelity to real sequences. Thus, we can infer that the synthetic data mimic real mRNAs very closely from a statistical standpoint. These results highlight the potential of advanced machine learning techniques, particularly those employing sophisticated embedding methods, in enhancing predictive performance in bioinformatics.

Our study compared the StructmRNA model with other machine learning models, focusing on training improvements, generalization gaps, and achieved losses. ANOVA tests showed a significant difference in training loss improvement (F = 8.76, p = 0.021), indicating varied effectiveness in reducing training loss. However, Public Test set loss improvements were similar across models. The generalization gap also varied significantly, showing differences in models' ability to generalize. Minimal training and Public Test set losses were 0.06 and 0.07, respectively, demonstrating effective loss minimization by the end of training (Fig. 8a, b). Moreover, Fig. 8c shows final training and Public Test set losses across various machine learning models, depicting improvements from initial to final values. Error bars indicate loss reduction over training iterations, highlighting each model's learning effectiveness. Blue and red markers denote training and Public Test set losses for nine models: StructmRNA variants (NCBI, OpenVaccine, GAN Data), Word2vec, ELMo, LSTM, CNN, VAE, and AGC. Bars indicate model learning and generalization: shorter bars mean less improvement, while longer bars show significant loss reduction. StructmRNA shows better training efficiency but inconsistent Public Test set performance. The generalization gap reveals varying overfitting or underfitting, highlighting complexities in NLP and machine learning model evaluation.



(a) Trained with Real RNA sequences

(b) Trained with 50% real and 50% synthetic RNA sequences

Fig. 7. Comparison of GAN model convergence in two scenarios.

During training, the GPU handles numerical computations, while the CPU manages data tasks and pre-training for StructmRNA. The detailed training loop and model complexity lead to higher CPU usage. For instance, pretraining BERT on the Open Vaccine dataset took 220 hours, used dual GPU T4s, had 70% CPU usage, and operated with a batch size of 64.

Discussion

We evaluated the StructmRNA model against various machine learning models based on three criteria: training and Public Test set loss improvement, the generalization gap (difference between training and Public Test set losses), and the lowest end losses achieved. ANOVA tests revealed a significant difference in training loss improvement among models, but no significant difference in Public Test set loss improvement (F-value = 0.98, *p* value = 0.354). The generalization gap showed significant variance (F-value = 5.52, *p* value = 0.047), indicating differences in models' ability to generalize from training to Public Test sets. The lowest end training and Public Test set losses were 0.06 and 0.07, respectively. These results suggest that while StructmRNA may improve training efficiency, this does not extend to Public Test set loss reduction. The generalization gap variance indicates different levels of overfitting or underfitting among the models. These findings help refine model selection and improvement for specific applications.

Previous studies show that combining neural network architectures improves RNA degradation prediction, like models integrating LSTMs and CNNs, which capture complex RNA features but struggle with efficiency

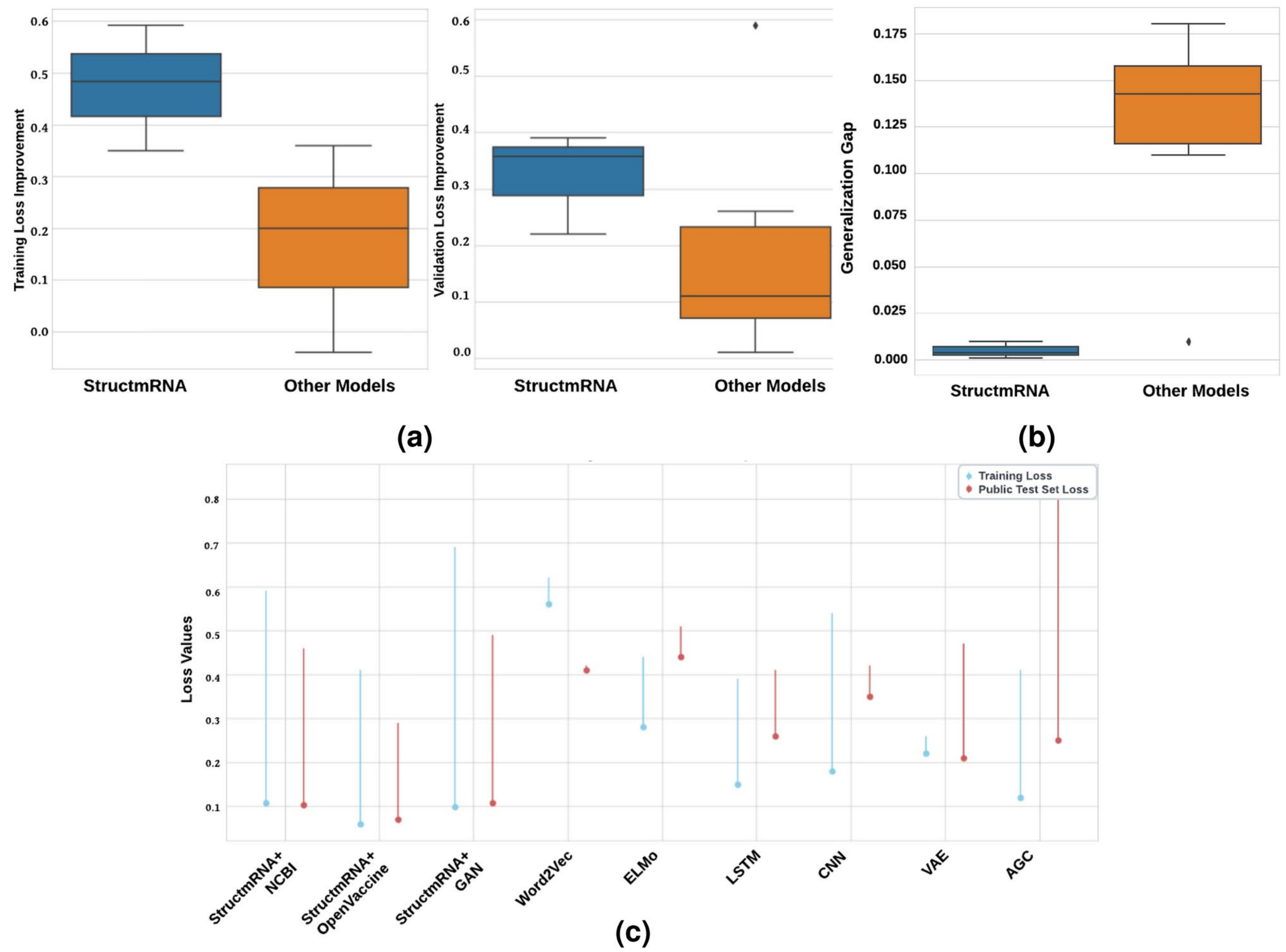


Fig. 8. Overview of training and Public Test set loss improvements and generalization gap. Overview: (**a**) Training and test loss improvements. (**b**) Generalization gap analysis: Box plot insights into model capabilities and overfitting. (**c**) Comparative visualization across models.

and overfitting. Genetic Algorithms optimize feature selection but miss RNA secondary structure details. Our StructmRNA model addresses these issues by using advanced embedding techniques and secondary structure information, achieving better accuracy with lower MCRMSE scores. This demonstrates the effectiveness of hybrid models and sophisticated embeddings in enhancing RNA degradation predictions.

We also compared training a BERT model on mixed synthetic (GAN-generated) and real RNA sequences from NCBI with training on only real NCBI data. Both scenarios converged within 30 epochs and performed similarly on RNA degradation tasks, indicating that BERT generalizes well across data types. Synthetic sequences exhibit high fidelity and can augment real datasets effectively. This highlights the need for optimizing synthetic and real data integration and refining GAN processes. Consistent performance underscores the importance of diverse metrics for evaluating synthetic data's impact. These results validate the use of synthetic data in bioinformatics and suggest that synthetic biology has potential in machine learning with challenging data acquisition.

Conclusion

The StructmRNA model significantly advances mRNA degradation prediction in bioinformatics. Our evaluation shows the superior performance and generalization of StructmRNA over various machine learning models, with notable improvements in training loss. However, Public Test set loss across different models remained similar, indicating comparable proficiency in mRNA degradation prediction. Exploring the generalization gap revealed differences in model adaptability from training to Public Test sets. Incorporating GAN-generated synthetic data into the training did not improve downstream performance but maintained model convergence within 30 epochs, highlighting the robustness of the BERT model and synthetic data's high fidelity to actual mRNA structures. This suggests the potential of integrating synthetic and real data without compromising predictive accuracy. StructmRNA addresses unique challenges by enhancing efficacy through regularization techniques, dropout layers, and optimized algorithms. Diverse Public Test sets ensure generalizability, and continuous evaluations highlight adaptability. Advanced preprocessing and regular updates maintain data quality and bias mitigation. Compatibility with existing bioinformatics tools and ongoing optimization efforts refine the training process, balancing complexity with performance.The mRNA degradation predictions of StructmRNA contributes to therapeutic research by identifying targets for mRNA-based drugs. This research highlights the role of advanced machine learning in bioinformatics and sets a new standard for degradation prediction. Synthetic and real data provide a model for future research, improving mRNA control in data-scarce fields for medical use.

Data availibility

The RNA degradation task used the OpenVaccine: COVID-19 mRNA Vaccine Degradation Prediction dataset from a Kaggle competition https://www.kaggle.com/competitions/stanford-covid-vaccine/data. The StructmRN A model, trained on NCBI Homo sapiens mRNAs, can be downloaded here (5 GB): https://www.kaggle.com/datasets/spnahali/sequences-ncbi-auto/data. The GAN dataset for DNAs used NCBI Homo sapiens DNAs and the GAN dataset is available at https://www.kaggle.com/datasets/spnahali/rna-seq-str. For more details, contact Sepideh Nahali at sepidnah@yorku.ca.

Received: 22 June 2024; Accepted: 21 October 2024
Published online: 29 October 2024

References

1. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120. https://doi.org/10.1093/bioinformatics/btab083 (2021).
2. Wayment-Steele, H. K. et al. Deep learning models for predicting RNA degradation via dual crowdsourcing. *Nat. Mach. Intell.* **4**, 1174–1184. https://doi.org/10.1038/s42256-022-00571-8 (2022).
3. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
4. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics, 2019).
5. Wang, J., Huang, J. X., Tu, X., Wang, J., Huang, A. J., Md Tahmid Rahman Laskar, and Bhuiyan, A. 2024. Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges. ACM Comput. Surv. 56, 7, Article 185 (July 2024), 33 pages. https://doi.org/10.1145/3648471
6. Wayment-Steele, H. K. et al. RNA secondary structure packages evaluated and improved by high-throughput experiments. *Nat. Methods* **19**, 1234–1242. https://doi.org/10.1038/s41592-022-01605-0 (2022).
7. Baden, L. R. et al. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *N. Engl. J. Med.* **384**, 403–416. https://doi.org/10.1056/NEJMoa2035389 (2021).
8. Banerjee, A. et al. Isolation, sequence, infectivity, and replication kinetics of severe acute respiratory syndrome coronavirus 2. *Emerg. Infect. Dis.* **26**, 2054–2063 (2020).
9. Coan, M., Haefliger, S., Ounzain, S. & Johnson, R. Targeting and engineering long non-coding RNAs for cancer therapy. *Nat. Rev. Genet.*[SPACE]https://doi.org/10.1038/s41576-024-00693-2 (2024).
10. Lai, H. et al. exoRBase 2.0: An atlas of mRNA, lncRNA and circRNA in extracellular vesicles from human biofluids. *Nucleic Acids Res.* **50**, D118–D128. https://doi.org/10.1093/nar/gkab1085 (2022).
11. Su, Y. et al. Plasma extracellular vesicle longRNAprofiles in the diagnosis and prediction of treatment response for breast cancer. *NPJ Breast Cancer* **7**, 154. https://doi.org/10.1038/s41523-021-00356-z (2021).
12. Li, S. et al. exoRBase: A database of circRNA, lncRNA and mRNA in human blood exosomes. *Nucleic Acids Res.* **46**, D106–D112. https://doi.org/10.1093/nar/gkx891 (2018).
13. Li, Y. et al. Extracellular vesicles longRNA sequencing reveals abundant mRNA, circRNA, and lncRNA in human blood as potential biomarkers for cancer diagnosis. *Clin. Chem.* **65**, 798–808. https://doi.org/10.1373/clinchem.2018.301291 (2019).
14. Liu, C. et al. Plasma extracellular vesicle long RNA in diagnosis and prediction in small cell lung cancer. *Cancers* **14**, 5493. https://doi.org/10.3390/cancers14225493 (2022).
15. Guo, T. et al. Plasma extracellular vesicle long RNAs have potential as biomarkers in early detection of colorectal cancer. *Front. Oncol.* **12**, 829230. https://doi.org/10.3389/fonc.2022.829230 (2022).
16. Li, Y. et al. Circulating EVs long RNA-based subtyping and deconvolution enable prediction of immunogenic signatures and clinical outcome for PDAC. *Mol. Ther. Nucleic Acids* **26**, 488–501. https://doi.org/10.1016/j.omtn.2021.08.017 (2021).
17. Li, Y. et al. EV-origin: Enumerating the tissue-cellular origin of circulating extracellular vesicles using exlr profile. *Comput. Struct. Biotechnol. J.* **18**, 2851–2859. https://doi.org/10.1016/j.csbj.2020.10.002 (2020).
18. Pederson, T. Review of "RNA: Life's indispensable molecule" by james e. darnell. *RNA*, **17**, 1771–1774. https://doi.org/10.1261/rna.2939711 (Cold Spring Harbor Laboratory Press, 2011).
19. Pyle, A. M. & Schlick, T. Challenges in RNA structural modeling and design. *J. Mol. Biol.* **428**, 733–735. https://doi.org/10.1016/j.jmb.2016.02.012 (2016).
20. Zhang, S. et al. Applications of transformer-based language models in bioinformatics: A survey. *Bioinform. Adv.* **3**, vbad001. https://doi.org/10.1093/bioadv/vbad001 (2023).
21. Iuchi, H. et al. Representation learning applications in biological sequence analysis. *Comput. Struct. Biotechnol. J.* **19**, 3198–3208. https://doi.org/10.1016/j.csbj.2021.05.039 (2021).
22. Jin, S., Zeng, X., Xia, F., Huang, W. & Liu, X. Application of deep learning methods in biological networks. *Brief. Bioinform.* **22**, 1902–1917. https://doi.org/10.1093/bib/bbaa043 (2020).
23. Akiyama, M. & Sakakibara, Y. Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. *NAR Genom. Bioinform.* **4**, lqac012. https://doi.org/10.1093/nargab/lqac012 (2022).
24. Shan, Y., Yang, J., Li, X., Zhong, X. & Chang, Y. GLAE: A graph-learnable auto-encoder for single-cell RNA-seq analysis. *Inf. Sci.* **621**, 88–103. https://doi.org/10.1016/j.ins.2022.11.049 (2023).
25. Xie, P., Zhuang, J., Tian, G. & Yang, J. Emvirus: An embedding-based neural framework for human-virus protein-protein interactions prediction. *Biosaf. Health* **5**, 152–158. https://doi.org/10.1016/j.bsheal.2023.04.003 (2023).
26. Eraslan, G., Avsec, Ž, Gagneur, J. & Theis, F. J. Deep learning: New computational modelling techniques for genomics. *Nat. Rev. Genet.* **20**, 389–403. https://doi.org/10.1038/s41576-019-0122-6 (2019).
27. Yi, H.-C., You, Z.-H., Huang, D.-S. & Kwoh, C. K. Graph representation learning in bioinformatics: Trends, methods and applications. *Brief. Bioinform.* **23**, bbab340. https://doi.org/10.1093/bib/bbab340 (2021).
28. Muzio, G., O'Bray, L. & Borgwardt, K. Biological network analysis with deep learning. *Brief. Bioinform.* **22**, 1515–1530. https://doi.org/10.1093/bib/bbaa257 (2020).
29. Wang, K., Hu, J. & Zhang, X. Identifying drug-target interactions through a combined graph attention mechanism and self-attention sequence embedding model. In *Advanced Intelligent Computing Technology and Applications* 246–257 (Springer Nature Singapore, 2023).
30. Cho, K. *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734 (2014).

31. Bowman, S. R. *et al.* Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, 10–21 (2016).
32. Zhang, Y. *et al.* Adversarial feature matching for text generation. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, vol. 70, 4006–4015 (2017).
33. Wang, L. & Zhou, Y. MRM-BERT: A novel deep neural network predictor of multiple RNA modifications by fusing BERT representation and sequence features. *RNA Biol.* **21**, 1–10. https://doi.org/10.1080/15476286.2024.2315384 (2024).
34. Zhang, L., Qin, X., Liu, M., Liu, G. & Ren, Y. BERT-m7G: A transformer architecture based on BERT and stacking ensemble to identify RNA n7-methylguanosine sites from sequence information. *Comput. Math. Methods Med.*[SPACE]https://doi.org/10.1155/2021/7764764 (2021).
35. Zhang, J., Fei, Y., Sun, L. & Zhang, Q. C. Advances and opportunities in RNA structure experimental determination and computational modeling. *Nat. Methods* **19**, 1193–1207. https://doi.org/10.1038/s41592-022-01623-y (2022).
36. Verbeke, R., Lentacker, I., De Smedt, S. C. & Dewitte, H. Three decades of messenger RNA vaccine development. *Nano Today* **28**, 100766 (2019).
37. Zhang, N. N. et al. A thermostable mRNA vaccine against covid-19. *Cell* **182**, 1271-1283.e16 (2020).
38. Wu, K. et al. Serum neutralizing activity elicited by mRNA-1273 vaccine. *N. Engl. J. Med.* **384**, 1468–1470 (2021).
39. Das, W. S. & et al. https://kaggle.com/competitions/stanford-covid-vaccine (2020).
40. Leppek, K. et al. Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. *Nat. Commun.Bold">13*, 1536. https://doi.org/10.1038/s41467-022-29272-w (2022).
41. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785 (ACM, 2016).
42. Wang, K., Zhu, Y., Huang, J. & Wang, X. A survey of complex network representation learning methods. In *Proceedings of the 2023 7th International Conference on High Performance Compilation, Computing and Communications*, HP3C '23, 160–168. https://doi.org/10.1145/3606043.3606066 (Association for Computing Machinery, New York, NY, USA, 2023).
43. Nishikawa, T., Lee, M. & Amau, M. New generative methods for single-cell transcriptome data in bulk RNA sequence deconvolution. *Sci. Rep.* **14**, 4156. https://doi.org/10.1038/s41598-024-54798-z (2024).
44. Marouf, M. et al. Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nat. Commun.* **11**, 166. https://doi.org/10.1038/s41467-019-14018-z (2020).
45. Murad, T. et al. Exploring the potential of gans in biological sequence analysis. *Biology* **12**, 854. https://doi.org/10.3390/biology12060854 (2023).
46. Lacan, A., Sebag, M. & Hanczar, B. GAN-based data augmentation for transcriptomics: Survey and comparative assessment. *Bioinformatics* **39**, i111–i120. https://doi.org/10.1093/bioinformatics/btad239 (2023).
47. Abdel-Basset, M., Moustafa, N. & Hawash, H. *Generative Adversarial Networks (GANs)*, 271–285 (2023).
48. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/.

Acknowledgements

We sincerely thank the editor and reviewers for their valuable feedback, which greatly improved the manuscript. We acknowledge support from NSERC (Grant RGPIN-2020-07157), the York Research Chairs program, and Compute Canada for computing resources.

Author contributions

S.N., L.S, and A.K conceived the results. S.N conducted and analysed them. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-77172-5.

Correspondence and requests for materials should be addressed to S.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.