Interpretable RNA Foundation Model from Unannotated Data for Highly Accurate RNA Structure and Function Predictions

Jiayang Chen[†1], Zhihang Hu[†1], Siqi Sun[*†2,3], Qingxiong Tan[1], Yixuan Wang[1,4], Qinze Yu[1,5], Licheng Zong[1], Liang Hong[1], Jin Xiao[1], Tao Shen[1,9], Irwin King[1], and Yu Li[*1,6,7,8,10]

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Research Institute of Intelligent Complex Systems, Fudan University, Shanghai, China
[3]Shanghai AI Laboratory, Shanghai, China
[4]Harbin Institute of Technology, China
[5]University of Electronic Science and Technology of China, China
[6]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA
[7]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, USA
[8]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[9]Zelixir Biotech, Shanghai, China
[10]The CUHK Shenzhen Research Institute, Hi-Tech Park, Nanshan, Shenzhen, China

Abstract

Non-coding RNA structure and function are essential to understanding various biological processes, such as cell signaling, gene expression, and post-transcriptional regulations. These are all among the core problems in the RNA field. With the rapid growth of sequencing technology, we have accumulated a massive amount of unannotated RNA sequences. On the other hand, expensive experimental observatory results in only limited numbers of annotated data and 3D structures. Hence, it is still challenging to design computational methods for predicting their structures and functions. The lack of annotated data and systematic study causes inferior performance. To resolve the issue, we propose a novel RNA foundation model (RNA-FM) to take advantage of all the 23 million non-coding RNA sequences through self-supervised learning. Within this approach, we discover that the pre-trained RNA-FM could infer sequential and evolutionary information of non-coding RNAs without using any labels. Furthermore, we demonstrate RNA-FM's effectiveness by applying it to the downstream secondary/3D structure prediction, SARS-CoV-2 genome structure and evolution prediction, protein-RNA binding preference modeling, and gene expression regulation modeling. The comprehensive experiments show that the proposed method improves the RNA structural and functional modelling results significantly and consistently. Despite only being trained with unlabelled data, RNA-FM can serve as the foundational model for the field.

Introduction

RNA plays an important role in performing various types of biological functions, such as cell signaling, gene expression, and post-transcriptional regulations [1–3]. Determination of RNA structure or type is also an essential part of RNA-based therapeutics, including mRNA vaccines, RNA interference and CRISPR-based therapeutics [4–6]. Among all RNA transcripts, about 5% served as mRNAs responsible for protein coding, while the substantial remaining portion is non-coding RNAs (ncRNAs) [7, 8]. Particularly, these ncRNAs must sustain specific structures to conduct corresponding biological functions. Different ncRNAs, such as small nuclear ribonucleoproteins (snRNPs), ribosomes, microRNAs, small nucleolar ribonucleoproteins (snoRNPs), long ncRNAs, and telomerase, also interact with proteins to form stable RNA-protein complexes to perform specific functions [1, 9]. Accurately modelling ncRNAs structures could help understand their performed functions and various biological processes. Despite a large number of ncRNA sequences, few of their structures and functions are known [10, 11]. Traditionally, RNA three-dimensional (3D) structures assessed by experimental approaches, including nuclear magnetic resonance, X-ray crystallography and cryogenic electron microscopy, are expensive and time-consuming. Therefore, computational approaches are developed and applied to bridge the gap.

[*]Corresponding Author. Email: liyu@cse.cuhk.edu.hk and siqisun@fudan.edu.cn.
[†]Equal first authorship.

Regarding the RNA structure prediction, most existing approaches focus on the RNA secondary structure prediction, which could further be divided into three categories: thermodynamic methods, alignment-based methods, and deep learning (DL)-based methods. Foundational works can be traced back to the 1980s, when thermodynamic parameters-based methods were proposed [12–14] to predict RNA secondary structures. Nowadays, combined with dynamic programming, these approaches such as Vienna RNA [15, 16], Mfold/UNAFold [17, 18], LinearFold [19], and RNAstructure [20, 21] are still widely used since they can incorporate a large volume of folding features (together with standard base pairs) to estimate the model parameters. Thermodynamic-based methods are later improved by considering parameters of local structures, including experimental parameters (such as RNAstructure [21], RNAfold [22], RNAshapes [23]) and machine learning parameters (such as ContextFold [24], CONTRAfold [25], CentroidFold [26]). Though many currently used approaches fall into this category, their overall performance seems to hit the plateau. These methods usually do not consider all base pairs obtained from tertiary interactions [14, 26, 27], which may miss essential information in their predictions.

To overcome the above problem, alignment-based methods built upon comparative sequence analysis are thus designed to determine the vital base pairs among homologous sequences [28]. With the help of sufficient homologous sequences and their alignments, these alignment-based methods achieve excellent performance in predictions [28, 29]. However, new issues arose from the limited number of known RNA families, given Rfam only contains several thousand RNA families. Since RNA is much less conserved, such property restricts the further improvement of alignment-based methods theoretically compared to the vast number of protein families. On the other hand, with more RNA data available, several deep learning approaches have been recently developed in the community to improve the accuracy of RNA secondary structure prediction. For example, SPOT-RNA [28], E2Efold [30], MXfold2 [31], and UFold [32] are shown to be able to improve the prediction accuracy significantly on different datasets. Nevertheless, the generalization capability of such DL-based methods still remains a problem, as the model architecture is explicitly designed for corresponding tasks and cannot generalize well to unknown RNA types [30].

In contrast to secondary structure prediction, the modelling of RNA 3D structure is still under-explored due to the lack of 3D structure data. In fact, computational optimization combined with deep learning methods may serve as an alternative way to solve the 3D problem. While there exist methods to optimize 3D structure with the minimum energy [33] given 2D information, deep learning methods could be utilized to solve the downgraded 2D problem (secondary structure, distance structure). For example, RNAcontact [34] could predict 3D closeness between base pairs by the deep residual neural networks. ARES, which consists of many processing layers, was proposed to score the predicted RNA structures [10]. Despite the above attempts so far, there are no end-to-end DL-based methods that could generate RNA 3D structure directly. The lack of annotated 3D data is one major obstacle.

In addition to RNA structure, understanding its function is also vital. In particular, predicting the interaction between RNAs and proteins [35, 36] could help to understand gene expression regulation. We could utilize either data from existing databases or data generated by biological experiments that classified RNAs into several functional groups for this application. With these extensive hand-crafted labels, DL-based approaches are then proposed to learn the underlying distribution of RNAs in different functional groups. Therefore, the corresponding functional group could be predicted given query RNA sequences. For instance, a deep discriminative neural network was developed to distinguish RNAs that can bind to specific RNA-binding proteins from the non-binding ones [36]. Furthermore, 1D CNN is designed to predict the protein expression levels regulated by human 5' UTRs from RNA sequence [37].

As the previous computational approaches rely heavily on the label information of RNA sequences, they all share the limitation of generalization. A model needs to be deliberately designed and tuned on specific tasks, yet challenging to transfer to other related studies. For instance, even E2Efold, a promising RNA secondary structure prediction model, performs well on dominant RNA types, including tRNAs and 5S rRNAs. Its performance on unknown RNA types degrades significantly. Also, models built upon RNA-protein interactions cannot gain accurate predictions on the UTR regulation. They are unfriendly to biological researchers because much effort will be needed to retrain specific deep learning based models.

Notice that virtually every method omits the millions of unannotated RNA sequences and only utilizes the small annotated dataset with at most 30K sequences. To overcome the above generalization limitation, we suggest taking advantage of the enormous unannotated RNA sequence data, which contains the evolutionary information of RNA sequences. We propose a novel RNA foundation model (RNA-FM), as shown in Figure 1, implying 'ONE-FOR-ALL', with which various RNA-related downstream applications can be conducted via replacing its predicting layers only. The model is trained in a self-supervised manner that differs from all the previous RNA DL-based approaches. To the best of our knowledge, we are the first that attempt to extract meaningful RNA representations from unlabeled data and improve multiple downstream applications simultaneously. Precisely,

the model is presented in two phases: pre-training and fine-tuning. Throughout the task-agnostic pre-training stage, the proposed RNA-FM is trained with 23 million ncRNA sequences from the RNAcentral database using self-supervised learning. Eventually, RNA-FM learns the sequential distribution and pattern that could potentially capture the underlying structural and functional information. Then, during the task-specific fine-tuning stage, the pre-trained RNA-FM model either generates sequence embeddings (features) that fit into downstream modules or is fine-tuned with a lightweight prediction layer. With the powerful representation learned from unannotated ncRNA data, RNA-FM significantly improves the performance across a broad range of RNA structure/function-related prediction applications with minor modifications to the model architectures. In terms of RNA secondary structure prediction, RNA-FM outperforms LinearFold by up to 30% and SPOT-RNA by up to 20% in terms of F1 score on cross-dataset validation. Even on the low-redundant dataset, which is rather challenging for previous methods, RNA-FM still outperforms SPOT-RNA by up to 7.5% and UFold by 4% in terms of F1 score. As for 3D closeness prediction, a single model built upon RNA-FM can even exceed an ensemble method with 100 models by 30% regarding the long-range top precision. In addition to working well on the benchmark datasets, RNA-FM can model the regulatory elements in the SARS-CoV-2 genome and potentially illustrates the evolution of the virus variants. Furthermore, the embedding from RNA-FM can achieve comparable performance with the *in vivo* secondary structure feature on protein-RNA interaction prediction, even though RNA-FM is trained on unannotated RNA sequences only. All of these performance improvements suggest that the proposed large-scale pre-trained foundation model has implicitly captured both structural and functional information from RNA sequences alone.

We summarize the main contributions of this work as follows:

- We propose the first RNA foundation model, RNA-FM, delivering rich representations for the ncRNA universe. In addition, we provide a web server such that the community can access the pre-trained model and its powerful representations. The server can be accessed with this link: https://proj.cse.cuhk.edu.hk/rnafm/ The code and weights are available at https://github.com/ml4bio/RNA-FM.
- RNA-FM can produce interpretable RNA representations, which contain evolutionary information. Such embeddings can be used to infer the evolutionary trend of lncRNAs and SARS-CoV-2 variants.
- The representations generated from RNA-FM can substantially improve the performance on multiple ncRNA structure/function prediction applications, with a desirable generalization property to the regulatory UTR regions of mRNAs and SARS-CoV-2.

Results

Learning from large-scale unlabeled non-coding RNA sequences. As shown in Figure 1, in order to take advantage of a massive amount of unlabeled ncRNA data and avoid relying on label information, we propose our RNA foundation model (RNA-FM) based on the BERT [38] language model architecture. It is built upon 12 transformer-based bidirectional encoder blocks and trained on 23 million sequences from the RNAcentral database in a self-supervised manner. After training, RNA-FM produces a $L \times 640$ embedding matrix for each RNA sequence with length L. These embeddings are expected to contain rich information within the ncRNA universe. We verified the effectiveness of RNA-FM on various applications. Firstly, to investigate what has been learned by the model and the physical meaning of the model outputs, we analyze the derived embeddings directly and examine how ncRNAs of similar function and structure gather in a 2-dimensional plane, resulting in the RNA Atlas in Figure 2. Also, the embedding from RNA-FM can be used to infer the long non-coding RNA (lncRNA) evolutionary trend, which suggests that the evolutionary information has been learned by our model implicitly. Furthermore, models using RNA-FM embeddings could improve over state-of-the-art approaches consistently on various structural-related and functional-related downstream prediction problems, including both SARS-CoV-2 study and gene expression regulation modeling.

RNA-FM learns multi-dimension biological information of RNA universe. To demystify what has been learned by the large-scale RNA-FM model from the million-scale data and the physical meaning of the model outputs, we take a closer look into RNA biological information contained in the RNA-FM embedding, including its structural/functional properties and evolutionary information. Such analysis shows the interpretability of the RNA-FM model.

RNA functions and structures vary across different RNA types, and RNA-FM is expected to encode these rich properties within our generated embeddings from pure RNA sequences. We build RNA Atlas by employing

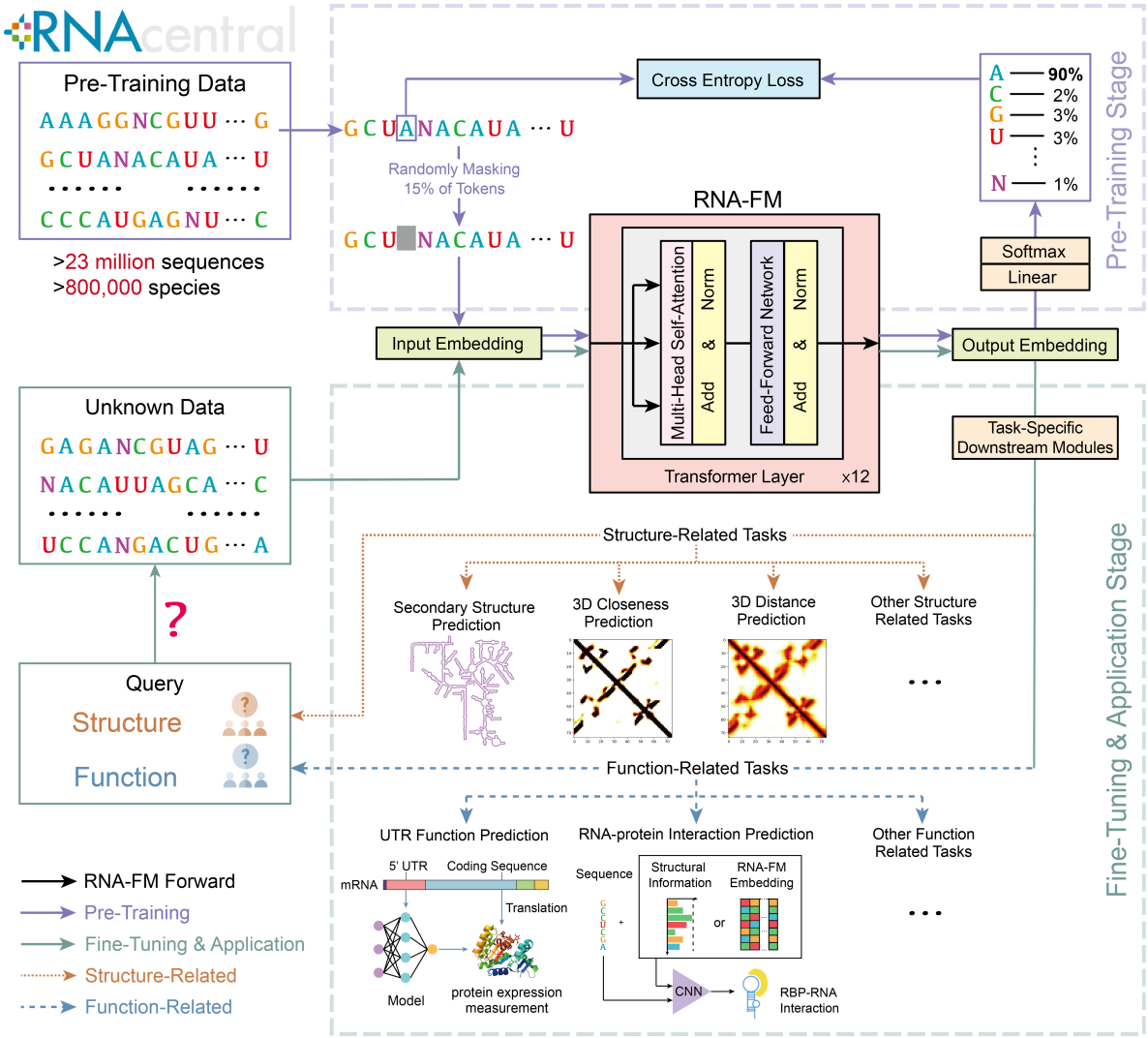


Figure 1: **Overview of RNA foundation model (RNA-FM) design and applications**. RNA-FM consists of 12 transformer layers. In the pre-training stage, the RNA-FM is trained with 23 million sequences from the RNAcentral database via self-supervised learning, *i.e.*, reconstructing the masked tokens from the sequence alone. Thus, we could obtain effective feature representations of the query RNA sequences without using any labels. In the fine-tuning stage, the representations from RNA-FM could significantly and consistently improve the performance on both structure-related and function-related applications with simple task-specific prediction modules.

the proposed encoder on the known RNA universe to analyze their patterns. Next, UMAP [39] is applied to reduce the dimension of embeddings and project them into a 2-dimensional plane. To compare and visualize how the learning process shapes the representations, we take both generated representations before and after model pre-training into consideration. Furthermore, random initialized RNA-FM (Random) and one-hot encoding (One-Hot) are also introduced for visualization purposes. The results are shown in Figure 2a. In the pre-trained RNA-FM embedding space (left), visual inspection reveals that RNA types are organized by structure and function properties with clear boundaries between clusters. In contrast, the projection of the Random model (middle) presents some extremely vague clustering structure; while projections of One-Hot encoding (right) are barely distinguishable with no apparent structure information, implying that RNA-FM has learned structural or functional information beyond their primary structure, such that instances with similar properties are grouped.

We take it a step further to discuss ncRNAs from different categorical views, such as housekeeping ncRNAs and regulatory ncRNAs, long ncRNAs (>200 nucleotides), and small ncRNAs (<=200 nucleotides). In Figures 2b(1) and b(3), RNA-FM well discriminates housekeeping and regulatory categories but struggles to deal with short and long ncRNAs. This might suggest that RNA-FM encoding emphasizes more on structural or functional similarity rather than length since RNAs with different lengths could share the same functions and RNAs with

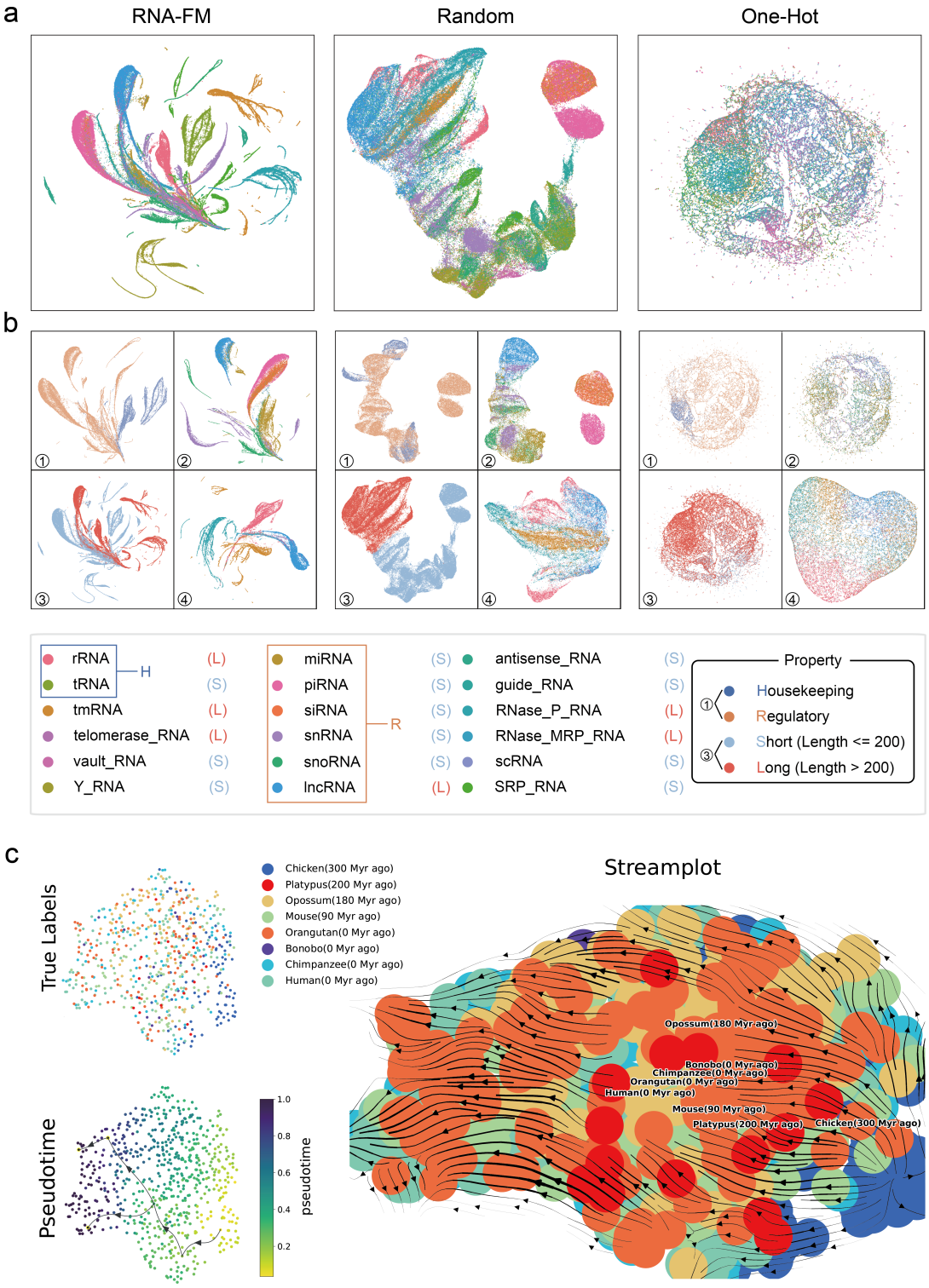


Figure 2: **RNA-FM encodes multi-scale patterns. a.** RNA Atlas with all ncRNA types in RNAcentral100 using different embedding ways. **b.(1)** Embedding projections of housekeeping RNA (rRNA, tRNA) and regulatory RNA (lncRNA, snoRNA, miRNA, siRNA, snRNA, piRNA). **b.(2)** Detailed distribution of the regulatory RNAs. **b.(3)** Embedding projections of long ncRNA (sequence length >200, rRNA, tmRNA, *etc.*) and small ncRNA (sequence length <200, tRNA, sncRNA, *etc.*). **b.(4)** Detailed distribution of long RNAs (most of the small RNAs have been analyzed in b.(2)). **c.** Trajectory inference of lncRNA evolutionary trend based on the RNA-FM embeddings.

similar lengths could differ significantly. In Figures 2b(2) and b(4), when we look closer into some sub-classes within a limited length range (less or greater than 200) in the RNA-FM part. These RNA embeddings aggregate or separate according to the similarity of their structures and functions; it proves again that RNA-FM establishes the RNA Atlas by recognizing the structures and functions of RNAs rather than their length.

RNA evolutionary information is also explored in our studies. We apply trajectory inference (pseudotemporal ordering) [40], which is commonly used in single-cell transcriptomics, to a subset of lncRNA with their RNA-FM embedding as input. RNAs in the subset can be classified according to different types of species. Here we obtain their evolutionary relationship from an evolutionary study of lncRNA repertoires and expression patterns [41]. We first generate RNA-FM embeddings for the lncRNA subset, then trajectory inference is carried out via VIA [42] and the stream-plot is shown in Figure 2c. We discover that although it is hard for RNA-FM to distinguish these RNAs into different species, the embeddings are able to present a roughly accurate evolutionary trend of different species corresponding to their ground-truth timeline. The result is surprising because we do not include such evolutionary features during training and only use the pure RNA sequences. This result testified that RNA-FM deeply mined the implicit genetic message and encoded the outputs with evolutionary information.

RNA-FM benefits both RNA secondary structure prediction and 3D modelling results. Structure understanding is always the key among various RNA-related applications since RNA structure usually determines its function. However, only a tiny fraction (<0.001%) of the structure-known ncRNAs has been determined by experiments [43] due to the high cost of wet-lab experiments and RNA structural instability. To tackle this problem, more and more computational approaches [28, 30–32] have been proposed for RNA structure prediction. We investigate RNA-FM's performance on several structure prediction tasks, including secondary structure prediction, 3D closeness prediction, and RNA map distance prediction. We also try to perform 3D reconstruction and prediction beyond the predicted secondary structures.

RNA secondary structure can be rapidly formed from its primary sequence by pairing bases with hydrogen bonds. Secondary structure is much more stable and more accessible in cells than its tertiary form, making it an essential role for the high-order structure prediction or even function prediction [43]. This section performs a comprehensive comparison of RNA-FM and other popular RNA secondary structure prediction methods, as well as a head-to-head comparison with one of the SOTA methods, UFold [32].

We conduct experiments on several benchmarks commonly used in E2Efold, SPOT-RNA and UFold. (1) RNAStralign [44], which consists of 37149 structures from 8 RNA types, is one of the most comprehensive collections of RNA structures in the field. (2) ArchiveII [45], which consists of 3975 RNA structures from 10 RNA types, is also a widely-used benchmark dataset for many classical RNA folding methods. (3) bpRNA-1m [28]. The dataset is preprocessed by removing sequence similarity with 80% sequence-identity cut-off and restricting their maximum sequence length below 500. The preprocessed dataset contains 13,419 sequences and is randomly split into 10,814 RNAs for training (TR0), 1300 for validation (VL0), and 1,305 for testing (TS0). For evaluation and testing, we take the usage of Ufold [32] data and make a fair comparison. The well-trained model is evaluated on ArchiveII600 (a subset with a length less than 600) and TS0.

Table 1 presents the accuracy of proposed RNA-FM and other advanced approaches [18, 19, 21, 22, 24, 25, 28, 30–32, 46, 47] to secondary structure prediction. RNA-FM outperforms all the other approaches concerning almost all metrics. In addition, RNA-FM far exceeds SPOT-RNA by a total of **22.8** points and **7.5** points on the ArchieveII600 and TS0, and distinctly higher than the SOTA (UFold) by a total of **3.4** points and **4.0** points on the ArchieveII600 and TS0, respectively, despite UFold also utilizes prior knowledge to model the probability of pairing. The superior performance demonstrates the advantages of underlying structural information encoded by RNA-FM.

Furthermore, we also conduct a head-to-head comparison of RNA-FM with UFold on ArchiveII600. Appendix Figure 7(a) shows the F1 score distribution across all samples in ArchieveII600, comparing the RNA-FM with UFold, corresponding to the y-axis and x-axis in the scatter plot, respectively. The RNA-FM matches or exceeds the UFold on **85.5%** of instances of all RNA types in the form of most points over the diagonal. We also explore the F1 score on different lengths of the input sequence, as shown in Appendix Figure 7(b). Regardless of the length of input RNA sequences, RNA-FM always outperforms UFold, especially when the RNA length is over 150, suggesting that our model better predicts the secondary structure of longer RNA sequences. Appendix Figure 8 presents the binary maps predicted by the model with a threshold of 0.5 and a graph view of the secondary structure predictions of two randomly selected examples. The probability maps from RNA-FM (second column) are more robust, less noisy, and much closer to the ground truth (first column) compared to those of UFold (third column). Regarding the graph-view converted from the binarized probability map, RNA-FM also generates secondary structures more similar to the ground truth than UFold.

Table 1: **RNA secondary structure prediction performance.** Our method beats the other 12 SOTA methods on the two datasets across all evaluation criteria except for being slightly behind Ufold on the recall score. RNA-FM is not specific to RNA secondary structure prediction. However, it has learned rich information about the secondary structure and done a much better job in RNA secondary structure prediction than other models.

Method	ArchiveII600 (3911)			bpRNA TS0 (1305)		
	Pre[a]	Rec	F1s	Pre	Rec	F1s
RNA-FM	**0.936**	**0.951**	**0.941**	**0.718**	0.713	**0.704**
UFold	0.890	0.926	0.905	0.607	**0.741**	0.654
E2Efold	0.738	0.665	0.690	0.140	0.129	0.130
LinearFold	0.641	0.617	0.621	0.561	0.581	0.550
Mfold	0.428	0.383	0.401	0.501	0.627	0.538
RNAstructure	0.563	0.615	0.585	0.494	0.622	0.533
RNAfold	0.565	0.627	0.592	0.494	0.631	0.536
CONTRAfold	0.607	0.679	0.638	0.528	0.655	0.567
SPOT-RNA	0.743	0.726	0.711	0.594	0.693	0.619
RNAsoft	0.665	0.594	0.622	0.497	0.626	0.535
MXfold2	0.788	0.760	0.768	0.519	0.646	0.558
Contextfold	0.873	0.821	0.842	0.529	0.607	0.546
Eternafold	0.667	0.622	0.636	0.516	0.666	0.563

[a] Pre, Rec, and F1s are the macro averages of the precision, recall and F1-score, respectively.

RNA 3D closeness indicates that arbitrary two bases have tertiary interaction if their distance is under a certain threshold, which originates from the "contact" concept in the protein field. Although secondary structure can reveal parts of the relationship between base pairs of RNA, it is merely a prior result and usually a constraint applied to the subsequent structure modelling. To obtain more precise structures, researchers propose many informative and challenging tasks for generating more strict constraints for downstream modelling methods. RNA 3D closeness utilizes a 2D matrix to represent pairwise tertiary inter-nucleotide interactions rather than the 2D flat relationship in secondary structure. The distance is defined as the minimal atomic distance of arbitrary bases, and the threshold is set as 8Å.

We select the benchmark datasets used by RNAcontact [34], which is constructed based on a set of non-redundant RNA 3D structures from Leontis and Zirbel (2012) (Version 3.99, 2019-11-06), containing 1786 entries with resolution < 4Å initially. Following preprocessing steps [34], we remove sequences with length $< 32nt$ or $> 1000nt$, with redundancy over 80% as well as with too few positive points (< 5). Finally, 221 sequences left are used for training (which we denoted as TR221), and 80 sequences for testing (denoted as TE80). The ground truth is computed from their PDB files following the steps above. The other features involved in the RNAcontact pipeline include the covariance of MSA and the secondary structure predicted by the PETfold [48] based on MSA. Appendix Figure 9 compares RNA 3D closeness prediction performance of RNAcontact and RNA-FMs with different initialization strategies and training schemes. Table 2 then presents the long-range top precisions of different models on the TE80 in detail. We focus on switching input representation features on the same ResNet32 architecture to achieve fair comparisons. Note that we only train ResNet32 once instead of averaging an ensemble of 100 models mentioned in RNAcontact [34]. A simple ResNet32 with RNA-FM embeddings achieves SOTA in all aspects and a great improvement over RNAcontact. Then we find that the long-range Top-L precision [49] improves **7** points when using our RNA-RM embeddings rather than using the covariance and PETfold prediction results. To pursue better performance on such a small dataset, we also adapt the transfer learning by initializing ResNet32 with the parameters pre-trained on bpRNA-1m in the above secondary prediction task. The transfer learning improves the performance significantly by another **20** points. Obviously, for the small-scale dataset, the pre-trained parameter is critical for both the backbone and downstream model. In addition, The long-range Top-L precision of the model with RNA-FM embedding is always higher than those with MSA covariances or the PETfold secondary structure, which indicates our embeddings from pure sequences present much more useful information than these features from MSA and further eliminate the time-consuming multiple sequence alignment generation step.

Appendix Figure 9(a) shows the long-range Top-L precision distribution across all samples in TE80, comparing the ResNet32 with different input features. The y-axis of the plot represents RNA-FM embedding with transfer

Table 2: **RNA 3D closeness prediction on the RNAcontact Test80 dataset (long-range top-precision).** The first row shows the results of RNAcontact with the sequence encoding as input (ensemble result of 100 models). The rest rows contain the results predicted by ResNet32 models with different feature inputs. The model with the RNA-FM embedding has already outperformed models with all the other features significantly (20% performance improvement on Top-L precision over RNAcontact), and its performance can be further boosted dramatically by transfer learning (33% performance improvement Top-L precision over RNAcontact). Unlike other features generated from MSA data, RNA-FM embedding is obtained from pure sequences, eliminating the time-consuming step of doing multiple sequence alignment.

Features	Source	Model	Long-Range Top Precision			
			L/10	L/5	L/2	L/1
Seq[a]	Seq	RNAcontact (100 ensemble)	0.48	0.45	0.40	0.33
Cov	MSA	ResNet32 (random)	0.57	0.54	0.45	0.34
Cov + SS[b]	MSA		0.62	0.61	0.54	0.46
RNA-FM[c]	Seq		<u>0.68</u>	<u>0.66</u>	<u>0.62</u>	<u>0.53</u>
RNA-FM	Seq	ResNet32 (transfer)	**0.88**	**0.85**	**0.79**	**0.66**

[a] *Seq* means sequence one-hot encoding
[b] *Cov* means MSA covariances; *SS* means secondary structure predicted by the PETfold based on MSA; + means a combination of features by a channel-wise concatenation.
[c] *RNA-FM* means the RNA-FM embeddings.

learning (*RNA-FM(TL)*), and the x-axis represents the combination of MSA covariance and secondary structure predicted by the PETfold as input (*Cov+SS*), which RNAcontact requires to generate from RNA MSA data. The RNA-FM with transfer learning matches or exceeds the MSA feature combination on **77.5%** of instances of all RNA types in the form of most points over the diagonal. We also explore the relationship between the Top-L precision and the input RNA sequence length, as shown in Appendix Figure 9(b). The RNA-FM embedding with transfer learning outperforms the MSA features across all sequence lengths.

Appendix Figure 9(c) presents the predicted probability maps of two randomly selected examples in the TS0. With the standalone RNA-FM embedding (*RNA-FM*, third column) as the input, the downstream model has already generated visualizations much closer to the ground truth (first column) than other features. Furthermore, we can perform far better than the other methods by applying above mentioned transfer learning (second column).

RNA distance map defines the distance of arbitrary bases in the primary sequence. In the past few years, more and more complex protein structure prediction tasks have been thoroughly studied. For instance, trRosetta [51] can predict the distance between two amino acids and the orientation formed by their atom planes, and Alphafold [52] can even directly predict the 3D structure of target proteins with high precision. However, 3D structure prediction in the RNA field is still an underdeveloped yet critical task. To ultimately approach this objective, we define a relatively new task for predicting this distance regression task, which can offer more information to downstream 3D folding methods than RNA secondary structure prediction [28] and 3D closeness prediction [34] mentioned above.

The dataset used for RNA distance prediction is the same as the benchmark used in the RNAcontact, as described in the previous paragraph. We generate distance maps for RNA sequences from their PDB files according to the minimal atomic distance of arbitrary bases. Then we limit the distance from 0 to 20 Å and regard the value over 20 as 20. Finally, we use 20 to normalize the distance values and obtain a normalized distance map with elements falling into $[0, 1]$. Table 3 summarizes the distance prediction performance of the model with different inputs on TE80. The U-Net with RNA-FM embeddings as input significantly outperforms sequences across different evaluation criteria with almost a 39% increase in R^2, a 39% increase in Pixel Accuracy (PA), a 41% increase in (PMCC), and 0.026 (42%) decrease in MSE. Furthermore, when simply combining RNA-FM embeddings with sequences, we already obtained a model with a slightly better MSE over approaches that take advantage of all features, such as secondary structure (*SS*) and MSA covariance (*Cov*). It suggests that our generated RNA-FM embedding contains the most explicitly helpful information for this task. Notice that RNA-FM is a pure single-sequence method, eliminating the time-consuming MSA searching step. Moreover, when combining RNA-FM embeddings with sequences and MSA covariances, we obtain the best model with the lowest MSE at 0.0319. Evaluated R^2 also achieves the highest value of 43% over the standalone sequence.

Detailed analysis is conducted between two models with different input features. One is the combination of sequences and RNA-FM embeddings (*RNA-FM+Seq*), and the other is the combination of sequences and the

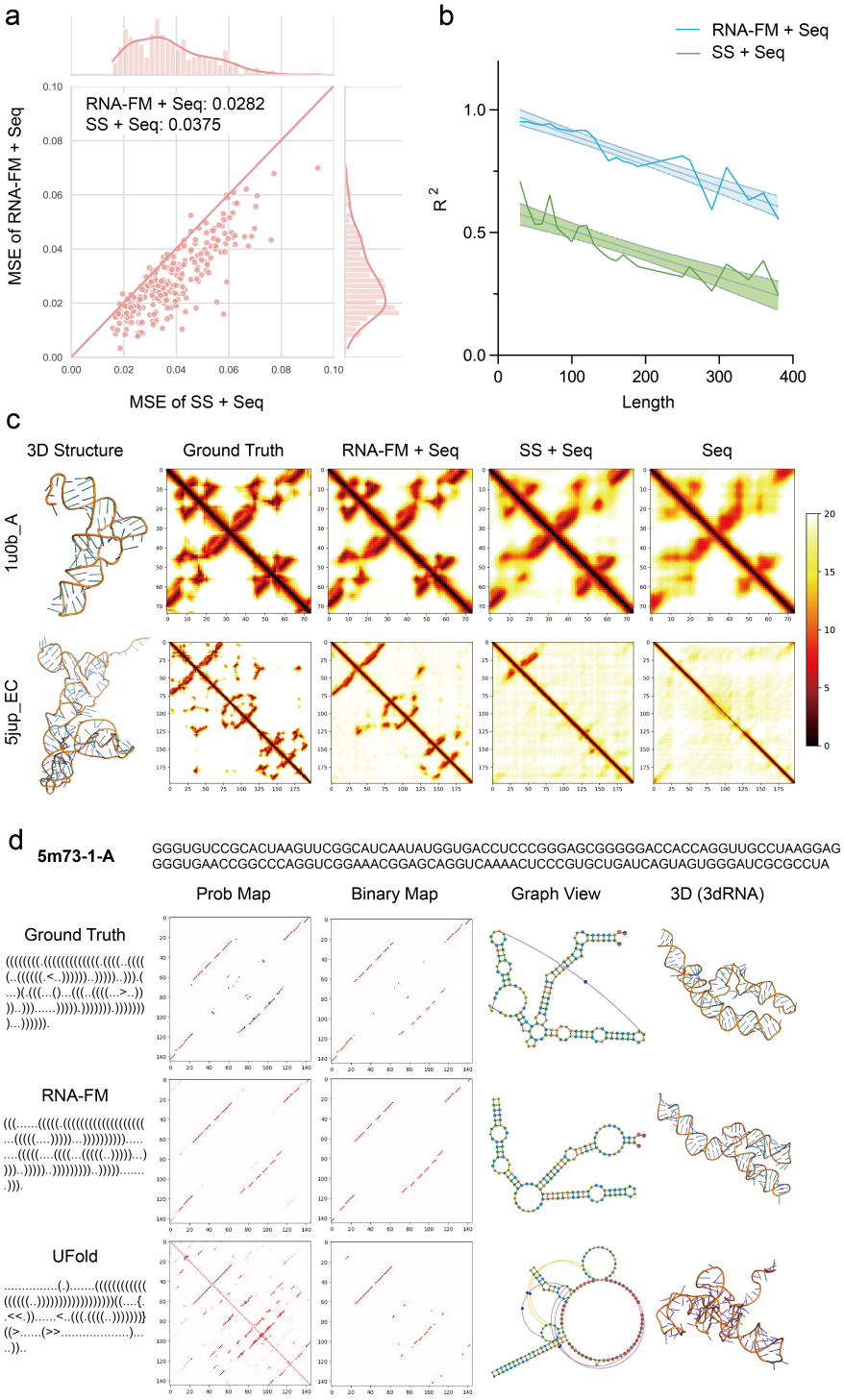


Figure 3: **RNA 3D distance prediction performance on the RNAcontact TE80 dataset and 3D reconstruction of RNA.** We use U-Net as the downstream base model with different input features. *Seq*: sequence one-hot encoding; *Cov* : MSA covariances; *SS*: secondary structure from E2Efold; *Emb*: embedding from RNA-FM; $+$: features combination via channel-wise concatenation. **a.** MSE Scatter plots with the performance of *RNA-FM+Seq* as y-axis and *SS+Seq* as x-axis. Each point represents an RNA structure. Almost all points are below the diagonal, which indicates that the RNA-FM embedding is superior over other features on nearly all instances. **b.** R^2 measurement as a function of RNA sequence lengths. *RNA-FM+Seq* significantly outperforms that of *SS+Seq* across all the lengths **c.** Probability maps of two randomly selected examples. Integrating RNA-FM embedding into input can significantly improve the model's performance. **d.** The probability maps and binary maps of an instance from PDB (5m73-1-A) are generated by different predictors. The graph views are obtained by jViz.Rna 4.0 [50]. Finally, the 3D structures are optimized and reconstructed by 3dRNA.

Table 3: **RNA 3D distance prediction performance on the RNAcontact TE80 dataset.** All the experiments are based on U-Net with different inputs. The model with standalone RNA-FM embedding can obtain a lower MSE than the model with sequence encoding and secondary structure information. When combined with sequence encoding, the RNA-FM embedding outperforms all the other feature combinations across different evaluation criteria. When combining sequence encoding, MSA covariances, and the RNA-FM embedding, we can reach the awe-inspiring performance of PMCC as high as 0.8313.

Features	MSE	R^2	PA(%)	PMCC
Seq[a]	0.0615	0.3652	45.77	0.4024
SS + Seq	0.0387	0.6875	81.45	0.7826
SS + Cov + Seq	0.0338	0.7821	85.44	0.8218
RNA-FM	0.0353	0.7542	84.62	0.8143
RNA-FM + Seq	0.0322	0.7824	86.13	0.8261
RNA-FM + Cov + Seq	**0.0319**	**0.7921**	**88.83**	**0.8313**

[a] *Seq* means sequences using one-hot encoding. *Cov* means MSA covariances. *SS* means secondary structures predicted by E2Efold. *RNA-FM* means RNA-FM embedding. + means a combination of features.

predicted secondary structures (*SS+Seq*). Figure 3(a) shows the MSE of these two models across all instances in TE80, with *RNA-FM+Seq* as the y-axis and *SS+Seq* as the x-axis. The *RNA-FM+Seq* is better than the *SS+Seq* on **94.2%** of instances in the form of most points below the diagonal. We also explore the relationship between the R^2 and the input RNA sequence length, as shown in Figure 3(b). When combined with sequence one-hot encoding, RNA-FM embedding outperforms the predicted secondary structure across all the RNA lengths.

As shown in Figure 3(c), we can see that our embedding feature enables the model to capture specific details of distance data while only providing sequential and secondary-structure data is not sufficient. The standalone sequential data does the worst in our experiment and only captures distance values on the diagonal.

RNA 3D reconstruction is the ultimate goal of RNA structure prediction. This section presents the results by combing RNA-FM with existing optimizing tools to obtain 3D approximates. Specifically, secondary structures are generated by the proposed RNA-FM and UFold. Then, their 3D structures can be optimized with RNA 3D modelling tools, including 3dRNA and FARFAR2 [53]. We employ 3dRNA here to optimize the 3D structure upon its secondary structure. By applying our model to the PDB dataset [28], we fine-tune the secondary structure predictor. The performance is shown in Supplementary Table 5. One of the examples in the PDB test set, namely TS1, is shown in Figure 3d. Notice that RNA-FM produces around 7.91 RMSD, which is significantly better than the results produced by UFold (25.70). Interestingly, even the ground truth secondary structure produces a higher RMSD (13.96), suggesting that the error may actually come from the 3D structure modelling process. Moreover, we also apply it to the DCS-PK element in the 5' UTR flanking region of the Zika virus (ZIKV) [54]. The DCS-PK is a pseudoknot found in the coding region [55], which helps enhance genome cyclization during replication. Due to the lack of ground truth for some 3D structures, RMSD for each prediction is unavailable, but RNA-FM produces more precise secondary structures than UFold for these targets, as shown in Appendix Figure 10.

To eliminate the deviation from the above 3D optimizing process, we also developed an end-to-end differentiable RNA 3D structures prediction model for evaluating the RNA-FM embedding by comparing them with the raw sequence inputs. As shown in Supplementary Table 6, on all the RNApuzzle structures, RNA-FM representations improve RNA 3D structure prediction greatly, with the average RMSD being around 4Å. It is consistent with all the above prediction tasks, suggesting that RNA-FM leads to more accurate RNA structure prediction.

RNA-FM facilities SARS-CoV-2 genome secondary structure and evolution study. COVID-19 has caused significant losses in properties and life in the past years, and detailed studies of the virus genome structure and its evolution are vital to prevent the next pandemic. We conduct such an investigation and apply RNA-FM to the whole genome of Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causal pathogen of the epidemic [57]. Firstly, we utilize the well-trained RNA-FM to predict the secondary structures of the key regulatory segments of the SARS-CoV-2 reference genome (Refseq accession number NC_045512.2). As shown in Figures 4(a) and (b), we sample 3'UTR, 5'UTR, and other segments from the entire genome with the length of 29870 based on the work of Cao *et al* [58]. Our model precisely predicts the majority of them. The results indicate that our model can effectively perform RNA secondary structure prediction task on an independent test set, with desirable generalization property. The predictions of the fragments (in black boxes) in 5'UTR are visualized

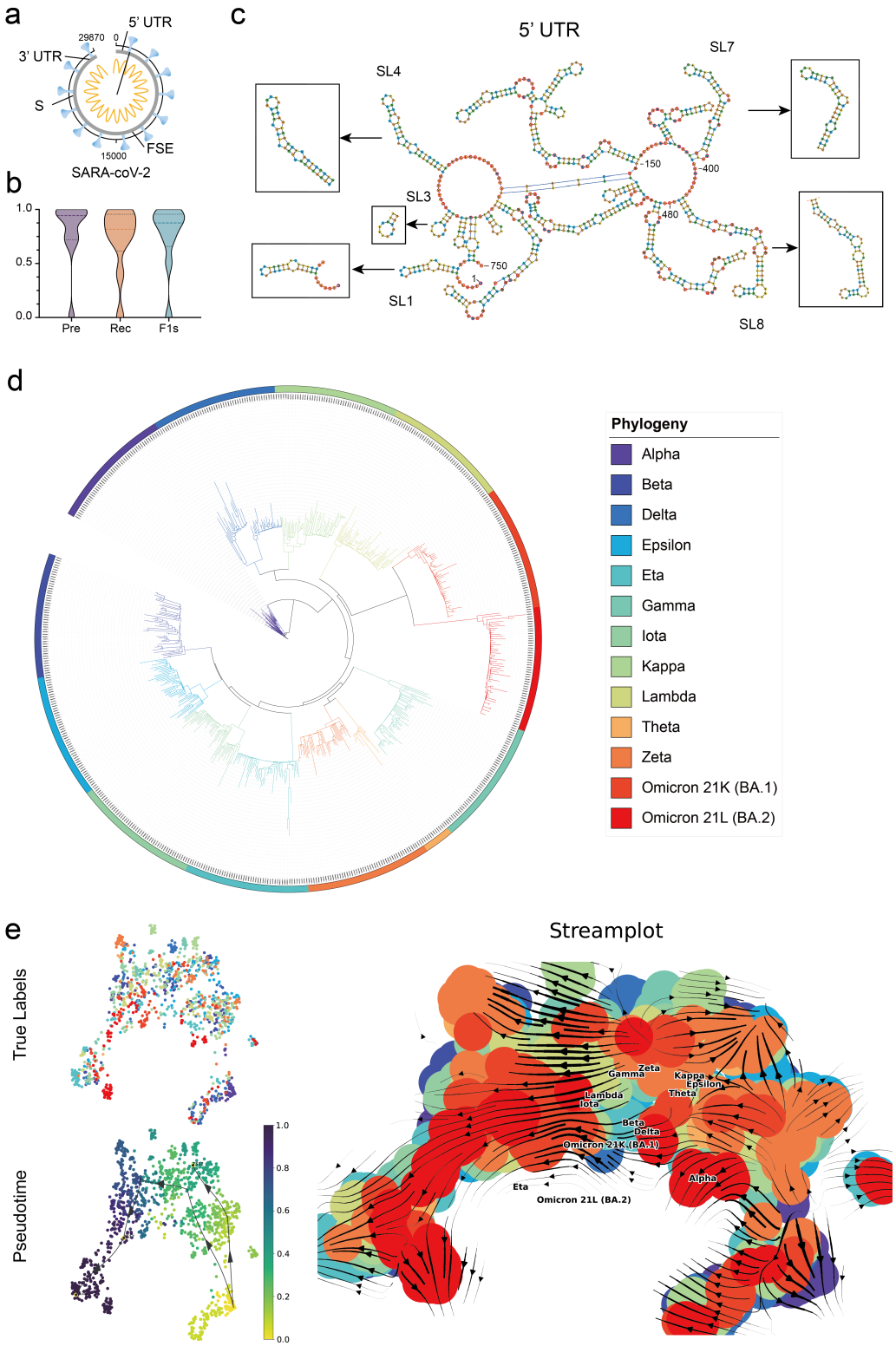


Figure 4: **RNA-FM predicts SARS-CoV-2 genome regulatory element secondary structures and the virus variant evolutionary trend.** **a.** Diagram of key regulatory segments sampled from the whole SARS-CoV-2 genome. **b.** Violin plots of the secondary structure prediction performance of segments mentioned above. RNA-FM can precisely predict the secondary structure of these segments. **c.** The visualization of RNA secondary structure predictions in 5'UTR. The prediction results are almost the same as the ground truth. **d.** The phylogenetic tree generated by FastME [56]. We treat it as the ground truth of the evolutionary trend of SARS-CoV-2, from Alpha types to Omicron variants. **e.** The trajectory inference of COVID-19 evolutionary trend with the RNA-FM genome-level embeddings. The results are highly consistent with the ground truth.

Table 4: **RBP-RNA binding prediction AUPRCs on the HeLa dataset with different RBPs.** CNN models with different input features are compared in this experiment. *RNA-FM+Seq* and *RealSS+Seq* outperform *Seq* on all RBPs except for METTL14. Although the *RealSS+Seq* achieves a higher mean AUPRCs than *RNA-FM+Seq*, the latter surpasses the former on nearly half of RBPs, which demonstrates the effectiveness of the RNA-FM embedding.

RBPs	Seq[a]	RNA-FM + Seq[b]	RealSS + Seq[c]
ELAVL1	0.946	**0.950**	0.938
UPF1	0.933	0.934	**0.956**
TIA1	0.917	**0.929**	0.925
METTL14	**0.889**	0.866	0.876
HNRNPC	0.874	**0.883**	0.873
CSTF2	0.860	0.867	**0.870**
TIAL1	0.853	0.852	**0.869**
U2AF65	0.852	**0.899**	0.886
eIF4AIII	0.818	0.814	**0.829**
WTAP	0.804	0.801	**0.843**
HNRNPU	0.798	**0.829**	0.821
PTBP1	0.793	0.799	**0.840**
PTBP1PTBP2	0.743	0.749	**0.775**
EIF4A3	0.738	0.751	**0.758**
U2AF2	0.712	0.705	**0.764**
METTL3	0.709	**0.750**	0.704
YTHDF2	0.610	**0.638**	0.629
Mean	0.815	0.824	**0.833**

[a] *Seq* means one-hot encoding of sequence (4 dims);
[b] *RNA-FM+Seq* combines the RNA-FM embedding (640 dims reduced to 1 dim) and the sequential one-hot encoding;
[c] *RealSS+Seq* means the combination of the structural score by the ic-SHAPE experiment with the sequence data.

in Figure 4(c), which are almost the same as the ground truth. Secondly, we explore the evolution of different COVID variants by applying RNA-FM to the whole genomes. Although RNA-FM is not initially designed for whole genome modelling, we assume that aggregation of the RNA-FM embedding extracted from fragments of the whole genome can still characterize the genome, benefiting the study of the virus genome evolution. The calculation of genome-level RNA-FM embedding can be seen in the Method section. As shown in Figure 4(e), the trajectory inference with the RNA-FM embedding of the virus genome is roughly in line with the phylogenetic tree generated by FastME [56], which is considered the ground truth (Figure 4(d)). The predictive evolution trend of COVID-19 begins with the Alpha type and ends up with the newest Omicron variant by April 2022, especially from Omicron 21K to Omicron 21L. Notice that RNA-FM is trained using merely the ncRNA sequences and unsupervised learning. We directly apply the trained model to the COVID-19 dataset without any fine-tuning or using any labelling information about the virus. It suggests that the regulatory elements of the virus genome could be vital for the virus variant evolution. Also, the RNA-FM framework can dig up core structure messages and evolutionary trend information of COVID-19 and its variants. Further development of the model has the potential to facilitate the research of COVID-19 and other pandemics.

RNA-FM carries secondary structure information for RNA-protein interaction modelling. Protein-RNA interactions are of vital importance in various cellular activities, including cell-signalling, post-transcriptional regulations, and protein synthesis [59]. We reproduce PrismNet [36], which includes *in vivo* RNA secondary structure profiles for RNA-protein interaction prediction. After that, we adopt RNA in the HeLa cell as the dataset for RNA binding protein prediction application and divide them into several sub-datasets according to different corresponding RBPs. The secondary structures of RNA are generated by icSHAPE (*in vivo* click selective 2'-hydroxyl acylation and profiling experiment) [60] in the HeLa cell environment. Then, we use our method to generate corresponding embeddings for all the sequences to replace the real secondary structures in PrismNet and make a comparison, evaluating the difference between the outputs from RNA-FM and the *in vivo* secondary

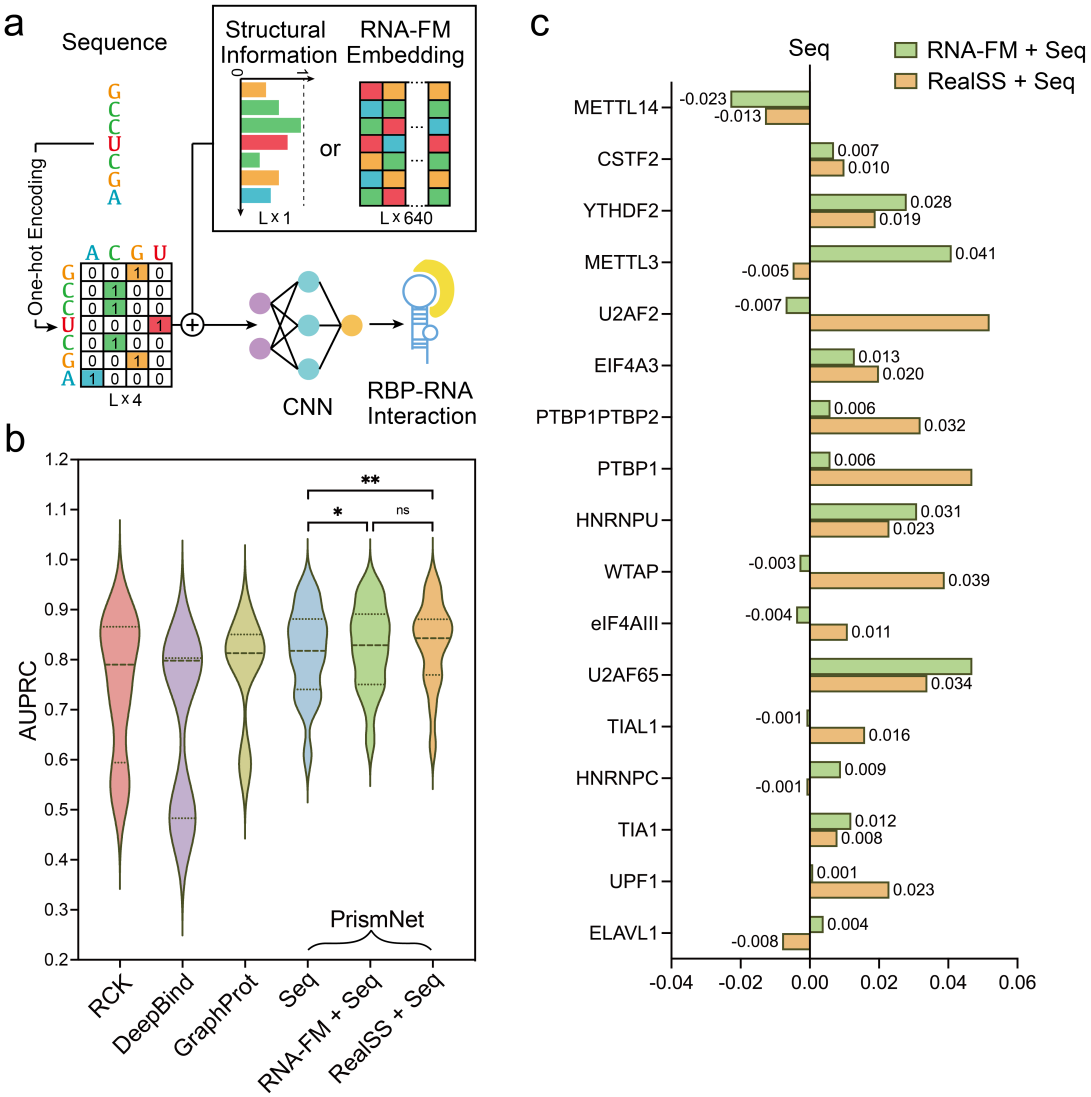


Figure 5: **RNA-FM embedding facilitates protein-RNA interaction prediction. a.** The deep learning framework for RBP-RNA interaction prediction in this experiment. The CNN model initially takes the combination of sequences and the structural scores offered by icSHAPE as input, while we substitute the experimental icSHAPE score with the RNA-FM embedding. **b.** Violin plots of AUPRCs of the model with different input features. Combining RNA-FM embedding with sequence encoding (*RNA-FM+Seq*) can improve the median and the quartiles of AUPRCs over sequence standalone, approaching the performance generated by the experiment-measured structural information (*RealSS+Seq*). **c.** Histogram plots of AUPRCs on different proteins with the *Seq* as the baseline, corresponding to the vertical line across the origin point. The *RNA-FM+Seq* outperforms the *Seq* in most cases and sometimes achieves even better performance than the *RealSS+Seq*, which shows the effectiveness of the RNA-FM features.

structure profiles.

For evaluation, we calculate AUPRCs on three streamlines (sequence only, sequence with real secondary structure, and sequence with RNA-FM) for comparison, as shown in Table 4 and Figure 5(c). RNA-FM embeddings with sequences achieve the best performance on nearly half of the subsets. Their performance is even comparable to the real secondary structure with sequences, suggesting that embeddings from RNA-FM provide sufficient information as real secondary structures. Furthermore, taking advantage of RNA-FM embeddings helps the original model improve performance over models only with sequential information. Figure 5(b) shows AUPRC violin plots for the three mentioned combinations as well as three other methods including RCK [61], DeepBind [62], GraphProt [63]. Our "RNA-FM+Seq" achieves close results compared to PrismNet with "Real SS+Seq". The RNA-protein interaction prediction results further illustrate that our embedding can learn sufficient information

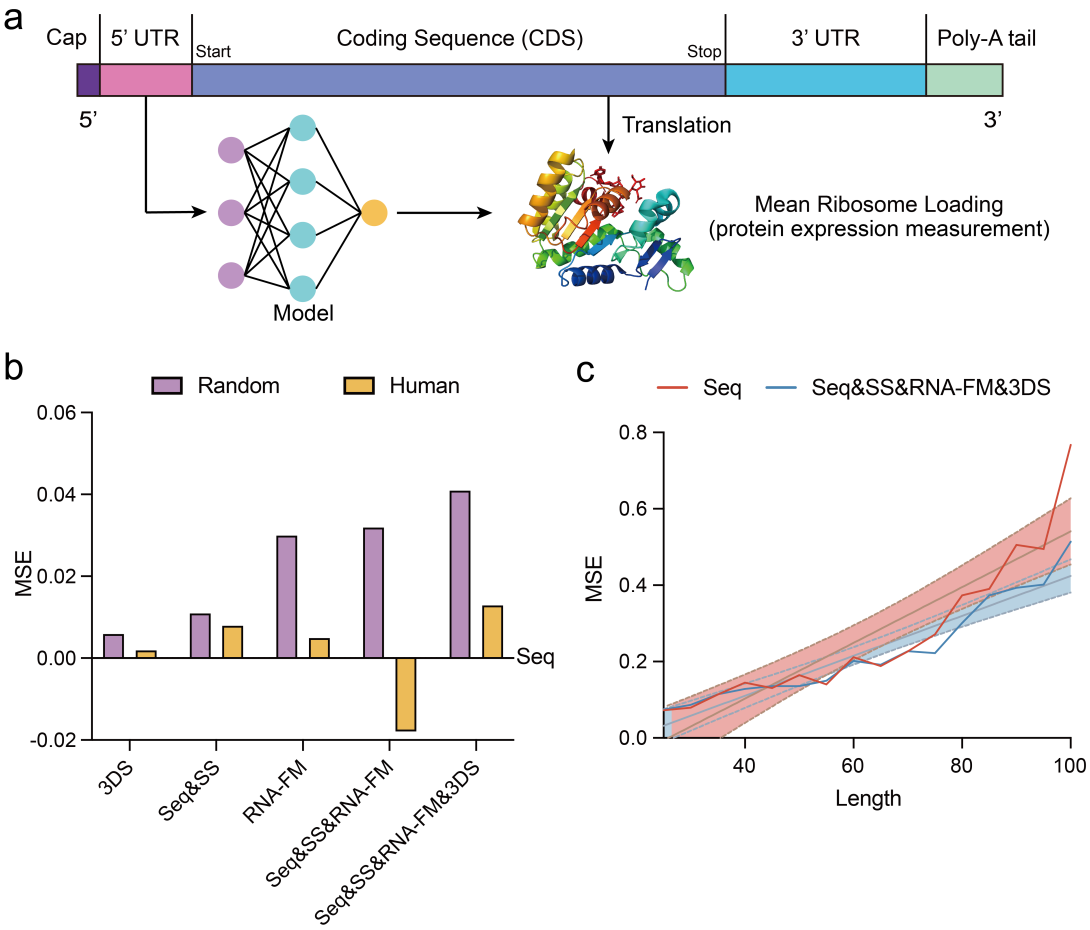


Figure 6: **5' UTR-based mean ribosome loading (MRL) prediction. a.** The deep learning framework for predicting MRL, which is a metric for evaluating the protein expression level regulated by UTR. **b.** Histogram plots of MSE on models of different input features with the one of *Seq* as the baseline, corresponding to the horizontal line across the origin point. The *RNA-FM* outperforms the *Seq* significantly on the Random Set. When combined with the secondary structure (SS) and 3D structure (3DS) information, which is predicted based on RNA-FM, our method can further improve the performance on the Human Set. **c.** R^2 as a function of RNA sequence lengths. The model with information based on RNA-FM significantly outperforms *Seq* model across all the lengths.

about secondary structures from raw sequences, which benefits the downstream functional prediction.

RNA-FM generalizes to mRNA untranslated region's function. We further assume that RNA-FM could directly benefit the gene expression regulation modeling, which is one of the ultimate goals in the related studies, because function partially depends on structures. The 5' untranslated region is the region of a messenger RNA (mRNA) located upstream of the initiation codon. This region is important for the translation regulation by different mechanisms in viruses, prokaryotes and eukaryotes. The sequence of 5' UTR is a primary determinant of mRNA translation efficiency, especially the concomitant coding sequence (CDS), which is responsible for target protein expression. Although RNA-FM is trained with ncRNAs and the 5'UTR is a part of an mRNA (not belong to ncRNAs), we test the versatility of RNA-FM to generalize it on handling implicit non-coding sequences of an mRNA and aid with modelling the relationship between UTR and target protein expression. With the assistance of massively parallel reporter assays and polysome profiling methods [37], which can measure the corresponding mean ribosome load (MRL) for each UTR, we can evaluate how a UTR regulates the target protein expression level of specific CDS by predicting the MRL of a UTR.

We utilize a large-scale synthetic Human 5'UTR library [37] as the dataset for the UTR function prediction task. The dataset consists of 83,919 5'UTRs of 75 different lengths and their corresponding MRLs. 7600 sequences are sampled equally at each length as a validation set, while the remainder is adopted for training. An

Table 5: **Mean ribosome loading (MRL) prediction performance on the Random7600 and Human7600 datasets.** Replacing the original input of MRL-CNN from sequence encoding (*Seq*) to RNA-FM embedding (*RNA-FM*) improves the performance.

Features	Random7600			Human7600		
	R^2	MAE	MSE	R^2	MAE	MSE
Seq[a]	0.860	0.371	0.277	0.814	0.375	0.269
Seq + SS[b]	0.866	0.369	0.266	0.820	0.370	0.261
RNA-FM[c]	0.876	0.360	0.247	0.816	0.377	0.264
3DS[d]	0.864	0.375	0.271	0.813	0.379	0.267
Seq + SS + RNA-FM	0.876	0.361	0.245	0.811	0.392	0.287
Seq + SS + 3DS + RNA-FM	**0.882**	**0.353**	**0.236**	**0.824**	**0.368**	**0.256**

[a] *Seq* means sequence encoding.
[b] *SS* means secondary structure, usually formatting an embedding (L$*$16) together with *Seq*.
[c] *RNA-FM* means RNA-FM embedding.
[d] *3DS* means embedding extracted from 3D structure prediction framework.

additional dataset consisting of 7600 real human 5'UTRs with the same length distribution provided by the library is used for validation to measure the generalization of models. We discover that model performances on the real human set are inferior to random set due to their data distribution difference, as illustrated in Table 5. For both datasets, the model with RNA-FM embeddings is better than the model with pure sequences. On the synthetic set, model based on RNA-FM embedding can achieve R^2 = 0.875 and MSE = 0.247. On the human set, it can achieve R^2 = 0.816 and MSE = 0.264. In addition, we assume that if we add more structure information to the model, we can further improve the modeling accuracy. So, we add the secondary structure information and 3D structure information, both predicted based on RNA-FM, into the model. As shown in Table 5 and 6(b), the prediction accuracy is indeed further improved. Besides, performance gains are consistent across all the lengths and contexts, as shown in Figure 6(c). This application further demonstrates the generalization of RNA-FM and its practical usage for real biological problems, even if the problem is not purely related to non-coding RNAs.

Discussion

To take advantage of the abundant unannotated RNA data, we propose an RNA foundation model trained on 23 million RNA sequences via self-supervised learning, which can be employed in both structural and functional downstream applications. Detailed analysis shows that RNA-FM also encodes the evolutionary information implicitly, which can be used to derive the evolutionary trend of lncRNAs and SARS-CoV-2 variants. Several further experiments, from structure prediction to gene expression regulation modeling, are conducted, and the results strongly prove the effectiveness of our pre-trained model. Particularly in structural-related experiments, models which include our RNA-FM embeddings can significantly improve the performance among various tasks spreading from simple to complex. When dealing with a complex task with a relatively large-scale dataset, it is more likely to achieve admirable performance by fine-tuning our RNA-FM and downstream modules together. In the case of simple task with small-scale datasets, it is better to utilize transfer learning to avoid over-fitting. On all accounts, our RNA-FM indeed encodes the RNA structural patterns and can offer explicit information useful for RNA structure predictions.

However, the improvement brought by RNA-FM in the functional tasks seems more slight compared with the gain in the structural tasks. The underlying reason may be the sequence distribution differences between these function-related applications and our pre-training dataset. Besides, the relation between the RNA structure and its function is too complicated to represent directly. Even though without colossal performance improvement, our embedding can still be beneficial for these downstream tasks. We aim to provide more impressive results regarding these functional-related tasks in the future.

References

[1] Miao, Z. & Westhof, E. Rna structure: advances and assessment of 3d structure prediction. *Annual review of biophysics* **46**, 483–503 (2017).

[2] Caprara, M. G. & Nilsen, T. W. Rna: versatility in form and function. *Nature structural biology* **7**, 831–833 (2000).

[3] Atkins, J. F., Gesteland, R. F. & Cech, T. Rna worlds: from life's origins to diversity in gene regulation (2011).

[4] Li, B., Niu, Y., Ji, W. & Dong, Y. Strategies for the crispr-based therapeutics. *Trends in pharmacological sciences* **41**, 55–65 (2020).

[5] Bora, R. S., Gupta, D., Mukkur, T. K. S. & Saini, K. S. Rna interference therapeutics for cancer: challenges and opportunities. *Molecular medicine reports* **6**, 9–15 (2012).

[6] Pardi, N., Hogan, M. J., Porter, F. W. & Weissman, D. mrna vaccines—a new era in vaccinology. *Nature reviews Drug discovery* **17**, 261–279 (2018).

[7] Chen, X. *et al.* Computational models for lncrna function prediction and functional similarity calculation. *Briefings in functional genomicss* **18**, 58–82 (2019).

[8] Wang, K. C. & Chang, H. Y. Molecular mechanisms of long noncoding rnas. *Molecular cell* **43**, 904–914 (2011).

[9] Cech, T. R. & Steitz, J. A. The noncoding rna revolution—trashing old rules to forge new ones. *Cell* **157**, 77–94 (2014).

[10] Townshend, R. J. *et al.* Geometric deep learning of rna structure. *Science* **373**, 1047–1051 (2021).

[11] Yao, R.-W., Wang, Y. & Chen, L.-L. Cellular functions of long noncoding rnas. *Nature cell biology* **21**, 542–551 (2019).

[12] Zuker, M. & Sankoff, D. Rna secondary structures and their prediction. *Bulletin of mathematical biology* **46**, 591–621 (1984).

[13] Waterman, M. S. & Smith, T. F. Rna secondary structure: A complete mathematical analysis. *Mathematical Biosciences* **42**, 257–266 (1978).

[14] Rivas, E. The four ingredients of single-sequence rna secondary structure prediction. a unifying perspective. *RNA biology* **10**, 1185–1196 (2013).

[15] Stadler, P. *et al.* Viennarna package 2.0. *Algorithms* (2011).

[16] Hofacker, I. L. *et al.* Fast folding and comparison of rna secondary structures. *Monatshefte für Chemie/Chemical Monthly* **125**, 167–188 (1994).

[17] Markham, N. R. & Zuker, M. Unafold. In *Bioinformatics*, 3–31 (Springer, 2008).

[18] Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research* **31**, 3406–3415 (2003).

[19] Huang, L. *et al.* Linearfold: linear-time approximate rna folding by 5'-to-3'dynamic programming and beam search. *Bioinformatics* **35**, i295–i304 (2019).

[20] Mathews, D. H., Andre, T. C., Kim, J., Turner, D. H. & Zuker, M. An updated recursive algorithm for rna secondary structure prediction with improved thermodynamic parameters (1998).

[21] Reuter, J. S. & Mathews, D. H. Rnastructure: software for rna secondary structure prediction and analysis. *BMC bioinformatics* **11**, 1–9 (2010).

[22] Lorenz, R. *et al.* Viennarna package 2.0. *Algorithms for molecular biology* **6**, 1–14 (2011).

[23] Janssen, S. & Giegerich, R. The rna shapes studio. *Bioinformatics* **31**, 423–425 (2015).

[24] Zakov, S., Goldberg, Y., Elhadad, M. & Ziv-Ukelson, M. Rich parameterization improves rna structure prediction. *Journal of Computational Biology* **18**, 1525–1542 (2011).

[25] Do, C. B., Woods, D. A. & Batzoglou, S. Contrafold: Rna secondary structure prediction without physics-based models. *Bioinformatics* **22**, e90–e98 (2006).

[26] Sato, K., Hamada, M., Asai, K. & Mituyama, T. Centroidfold: a web server for rna secondary structure prediction. *Nucleic acids research* **37**, W277–W280 (2009).

[27] Nowakowski, J. & Tinoco Jr, I. Rna structure and stability. In *Seminars in virology*, vol. 8, 153–165 (Elsevier, 1997).

[28] Singh, J., Hanson, J., Paliwal, K. & Zhou, Y. Rna secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature communications* **10**, 1–13 (2019).

[29] Gutell, R. R., Lee, J. C. & Cannone, J. J. The accuracy of ribosomal rna comparative structure models. *Current opinion in structural biology* **12**, 301–310 (2002).

[30] Chen, X., Li, Y., Umarov, R., Gao, X. & Song, L. Rna secondary structure prediction by learning unrolled algorithms. *arXiv preprint arXiv:2002.05810* (2020).

[31] Sato, K., Akiyama, M. & Sakakibara, Y. Rna secondary structure prediction using deep learning with thermodynamic integration. *Nature communications* **12**, 1–9 (2021).

[32] Fu, L. *et al.* Ufold: fast and accurate rna secondary structure prediction with deep learning. *bioRxiv* 2020–08 (2021).

[33] Xiong, P., Wu, R., Zhan, J. & Zhou, Y. Pairing a high-resolution statistical potential with a nucleobase-centric sampling algorithm for improving rna model refinement. *Nature Communications* **12**, 1–11 (2021).

[34] Sun, S., Wang, W., Peng, Z. & Yang, J. Rna inter-nucleotide 3d closeness prediction by deep residual neural networks. *Bioinformatics* **37**, 1093–1098 (2021).

[35] Lam, J. H. *et al.* A deep learning framework to predict binding preference of rna constituents on protein surface. *Nature communications* **10**, 1–13 (2019).

[36] Sun, L. *et al.* Predicting dynamic cellular protein–rna interactions by deep learning using in vivo rna structures. *Cell research* **31**, 495–516 (2021).

[37] Sample, P. J. *et al.* Human 5' utr design and variant effect prediction from a massively parallel translation assay. *Nature biotechnology* **37**, 803–809 (2019).

[38] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[39] McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[40] Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nature biotechnology* **37**, 547–554 (2019).

[41] Necsulea, A. *et al.* The evolution of lncrna repertoires and expression patterns in tetrapods. *Nature* **505**, 635–640 (2014).

[42] Stassen, S. V., Yip, G. G., Wong, K. K., Ho, J. W. & Tsia, K. K. Generalized and scalable trajectory inference in single-cell omics data with via. *Nature communications* **12**, 1–18 (2021).

[43] Zhao, Q. *et al.* Review of machine learning methods for rna secondary structure prediction. *PLoS computational biology* **17**, e1009291 (2021).

[44] Tan, Z., Fu, Y., Sharma, G. & Mathews, D. H. Turbofold ii: Rna structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic acids research* **45**, 11570–11581 (2017).

[45] Sloma, M. F. & Mathews, D. H. Exact calculation of loop formation probability identifies folding motifs in rna secondary structures. *RNA* **22**, 1808–1818 (2016).

[46] Andronescu, M., Aguirre-Hernandez, R., Condon, A. & Hoos, H. H. Rnasoft: a suite of rna secondary structure prediction and design software tools. *Nucleic acids research* **31**, 3416–3422 (2003).

[47] Wayment-Steele, H. K., Kladwang, W., Participants, E. & Das, R. Rna secondary structure packages ranked and improved by high-throughput experiments. *BioRxiv* (2020).

[48] Seemann, S. E., Gorodkin, J. & Backofen, R. Unifying evolutionary and thermodynamic information for rna folding of multiple alignments. *Nucleic Acids Research* **36**, 6355–6362 (2008).

[49] Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS computational biology* **13**, e1005324 (2017).

[50] Shabash, B. & Wiese, K. C. Numerical integration methods and layout improvements in the context of dynamic rna visualization. *BMC bioinformatics* **18**, 1–18 (2017).

[51] Yang, J. *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences* **117**, 1496–1503 (2020).

[52] AlQuraishi, M. Alphafold at casp13. *Bioinformatics* **35**, 4862–4865 (2019).

[53] Watkins, A. M., Rangan, R. & Das, R. Farfar2: improved de novo rosetta prediction of complex global rna folds. *Structure* **28**, 963–976 (2020).

[54] Li, P. *et al.* Integrative analysis of zika virus genome rna structure reveals critical determinants of viral infectivity. *Cell host & microbe* **24**, 875–886 (2018).

[55] Liu, Z.-Y. *et al.* Novel cis-acting element within the capsid-coding region enhances flavivirus viral-rna replication by regulating genome cyclization. *Journal of virology* **87**, 6804–6818 (2013).

[56] Desper, R. & Gascuel, O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. In *International Workshop on Algorithms in Bioinformatics*, 357–374 (Springer, 2002).

[57] Wu, F. *et al.* A new coronavirus associated with human respiratory disease in china. *Nature* **579**, 265–269 (2020).

[58] Cao, C. *et al.* The architecture of the sars-cov-2 rna genome inside virion. *Nature communications* **12**, 1–14 (2021).

[59] Wei, J., Chen, S., Zong, L., Gao, X. & Li, Y. Protein-rna interaction prediction with deep learning: structure matters. *Briefings in Bioinformatics* (2021). URL https://doi.org/10.1093/bib/bbab540.

[60] Spitale, R. C. *et al.* Structural imprints in vivo decode rna regulatory mechanisms. *Nature* **519**, 486–490 (2015).

[61] Orenstein, Y., Wang, Y. & Berger, B. Rck: accurate and efficient inference of sequence-and structure-based protein–rna binding models from rnacompete data. *Bioinformatics* **32**, i351–i359 (2016).

[62] Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology* **33**, 831–838 (2015).

[63] Maticzka, D., Lange, S. J., Costa, F. & Backofen, R. Graphprot: modeling binding preferences of rna-binding proteins. *Genome biology* **15**, 1–18 (2014).

[64] Vaswani, A. *et al.* Attention is all you need. In *Advances in neural information processing systems*, 5998–6008 (2017).

[65] Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[66] Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118** (2021).

[67] Rnacentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic acids research* **49**, D212–D220 (2021).

[68] Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

Methods

Overview of RNA-FM. In order to advance the development of RNA studies, we aim to build a unified foundation model in providing rich and meaningful representations inferred from standalone sequential information. We expect such representations to significantly boost various downstream tasks' performances when sufficient annotated data are unavailable. Inspired by recent great success in natural language processing, computer vision, and bioinformatics [64][65][66], we explore the possibilities of a general transformer architecture in RNA-related studies. Thus, our framework is built upon the bidirectional transformer language model proposed in BERT [38], followed by the unsupervised training scheme. We named our framework RNA-FM, suggesting a foundational model for all the RNA-related studies.

This section first illustrates how we construct the large-scale ncRNA dataset, followed by model and training details. Next, we investigate a general strategy for several RNA function-related or structure-related downstream applications.

Large-scale pre-training dataset. Our large-scale dataset used for pre-training phase is collected from RNAcentral [67], the to-date largest dataset of the ncRNA. This dataset is indeed a comprehensive ncRNA sequences collection, representing all the ncRNA types from a broad range of organisms. It combines ncRNA sequences across 47 different databases, adding up to around 27 million RNA sequences in total.

We then pre-process all ncRNA sequences by replacing 'T's with 'U's since they are both complementary to adenine and similar in structure ('T's used for representing thymine in DNA while 'U's stands for uracil in RNA.). This results in a dataset involving 4 main types of bases (16 counted types of combination in total, 'A', 'C', 'G', 'U', 'R', 'Y', 'K', 'M', 'S', 'W', 'B', 'D', 'H', 'V', 'N', '-'). Moreover, to reduce redundancy without hurting the capacity of our dataset (reserve sequences as more as possible), we eliminate identical sequences by applying cd-hit-est [68] with a cut-off at 100%.

After the above pre-processing steps, a final large-scale dataset consisting of 23.7 million ncRNA sequences is obtained. We named it RNAcentral100, which will be used to train our RNA foundation model in a self-supervised manner. Please refer to Supplementary Table.1-3 for the distribution of nucleotides and sequence lengths.

RNA foundation model training details. Our RNA-FM's framework is a stack of 12 transformer encoder blocks proposed in BERT [38][64]. Each encoder block [38] consists of a 640 hidden size feed-forward layer and a 20 multi-heads self-attention layer. Layer normalization [38] and residual connections are applied before and after every block, respectively.

For a RNA sequence with length L, RNA-FM takes raw sequential tokens as input, and an embedding layer maps each nucleotide token into a 640-dimensional vector, thus resulting in an $L \times 640$ embedding matrix. The embedding matrix then proceeds through each encoder block, which includes multi-head self-attention modules and feed-forward layers. The output tensors from encoder blocks have the exact same size with the input, and a final Softmax layer is concatenated above to predict corresponding tokens including our selected 16 nucleotides and 4 specific functional identifiers.

During the pre-training phase, we followed a self-supervised training manner in BERT [38]. Around 15% of nucleotide tokens are randomly replaced with a special mask token. (If the i-th token is chosen, we replace the i-th token with (1) the [MASK] token 80% of the time (2) a random token 10% of the time (3) the unchanged i-th token 10% of the time). We then train the model with masked language modeling (MLM) [38] by predicting the original masked token with cross-entropy loss. Such a training strategy can be formulated as an objective function as follow:

$$\mathcal{L}_{MLM} = \mathbb{E}_{x \sim \mathcal{X}} \mathbb{E}_{x_{\mathcal{M}} \sim x} \sum_{i \in \mathcal{M}} - \log p(x_i | x_{/\mathcal{M}}). \tag{1}$$

Here in the above equation, a set of indices $\mathcal{M}$ are randomly sampled from each of the input sequences x (15% among the whole sequence), and masked by replacing the true token at each index i to some other mask tokens. Next, for each masked token, when the masked sequence ($x_{/\mathcal{M}}$) is given as context, our adopted objective function will minimize the negative log-likelihood of the corresponding true nucleotide x_i. Our adopted objective function can capture dependencies between masked proportion and the remaining parts of the input sequence, which enables accurate predictions for masked positions. Thus, RNA-FM trained via Eq. (1) drives the network to gain deep understanding and rich representation of each sequential token.

In summary, we trained RNA-FM on eight A100 GPUs of 80 GB memories for one month. We adopted an inverse square root learning rate schedule for the training of the neural network, with a 0.0001 base learning rate,

a 0.01 weight decay, and 10000 warm-up steps. Similar to the previous studies, we also set the maximum length of the input sequences as 1024 to reduce the memory consumption and increase the batch size to speed up the training of the neural network.

RNA foundation model-generated embeddings analysis. After the training stage, we can get the embeddings for all the input RNA sequences generated by our RNA foundation model. In order to test if these embeddings have fully mined the structural and functional information belongs to RNA, we build RNA Atlas by subsample RNAcentral100 with a maximum of 10,000 samples per each RNA types. Each instance from different RNA families can be represented by a 640 dimensional vector by averaging across its RNA-FM embedding with shape of $L * 640$ at each position in its sequence, where L denotes the sequence length. Then, we apply UMAP [39] to reduce these 640 dimensional vectors into 2 dimensional ones, and project them on a plane. Furthermore, we explore if the embeddings contain evolutionary information of RNA by applying trajectory inference, which is implemented by VIA [42]. We take embeddings of lncRNA as inputs, and get the stream-plot of RNA evolutionary trend.

Downstream training strategies. Once the pre-trained RNA-FM model encoded with RNA's informative structural and functional patterns is obtained, we can integrate it into various downstream applications via two schemes: feature-based training and fine-tuning. For the feature-based scheme, we freeze the parameters inside RNA-FM and feed them to downstream models by switching the output module. In contrast, the fine-tuning scheme will require training RNA-FM together with downstream modules rather than freezing parameters inside RNA-FM, which usually performs better than the feature-based training scheme.

Our proposed RNA-FM can effectively tackle both the RNA structure prediction and function modeling. We first investigate structure-related applications, including RNA secondary structure prediction, RNA 3D closeness prediction, and RNA distance prediction. Additionally, we investigate the downstream function-related tasks, including COVID-19 virus modeling, RNA-protein interaction prediction, and gene expression regulation modeling. We give more details about how to set up our methodology to perform in the major downstream applications.

RNA second-order structure modeling. RNA high-order structures can usually be represented by a 2D matrix. We would focus on three related studies to highlight the capability of RNA-FM: secondary structure, contact map, and distance map prediction. Secondary structure reflects its hydrogen bonds in primary sequence. While contact and distance maps concentrate more on pairwise tertiary inter-nucleotide interactions. Here, we adopt a simple 2D ResNet as a unified downstream prediction module for all the structure prediction tasks rather than creating an elaborately-designed framework for each sub-task. Similar to ESM-1b [66], our deep residual network consists of 32 blocks, where each contains two convolution layers with a filter size of 64. The input of the ResNet32 is the outer concatenation of output embeddings obtained from query sequences. This module is utilized across all the RNA structure prediction tasks unless we specify (See Supplementary for more details of ResNet32). We show that RNA-FM can achieve significantly better performance than the state-of-the-art models with such simple downstream modules.

End-to-end RNA 3D structure prediction. Similar to AlphaFold2 in protein research, we aim to first establish such an end-to-end differentiable model for RNA 3D structure prediction. The model takes advantage of 4 Evoformers as its backbone, and we stack an equivariant graph transformer (EGNN) on top as a 3D atom coordinate predictor. The data are collected via the up-to-date Protein Data Bank, including raw RNA sequences with corresponding 3D structures. 1036 sequences are used for training, and the left 100 are used for validation and testing purposes. We utilize distance RMSD as our objective for training, and we optimize the model for more than 10000 steps to get an initial version. To validate the effectiveness of our RNA-FM embeddings, we compare it with the pure RNA sequences as the inputs of the 3D framework.

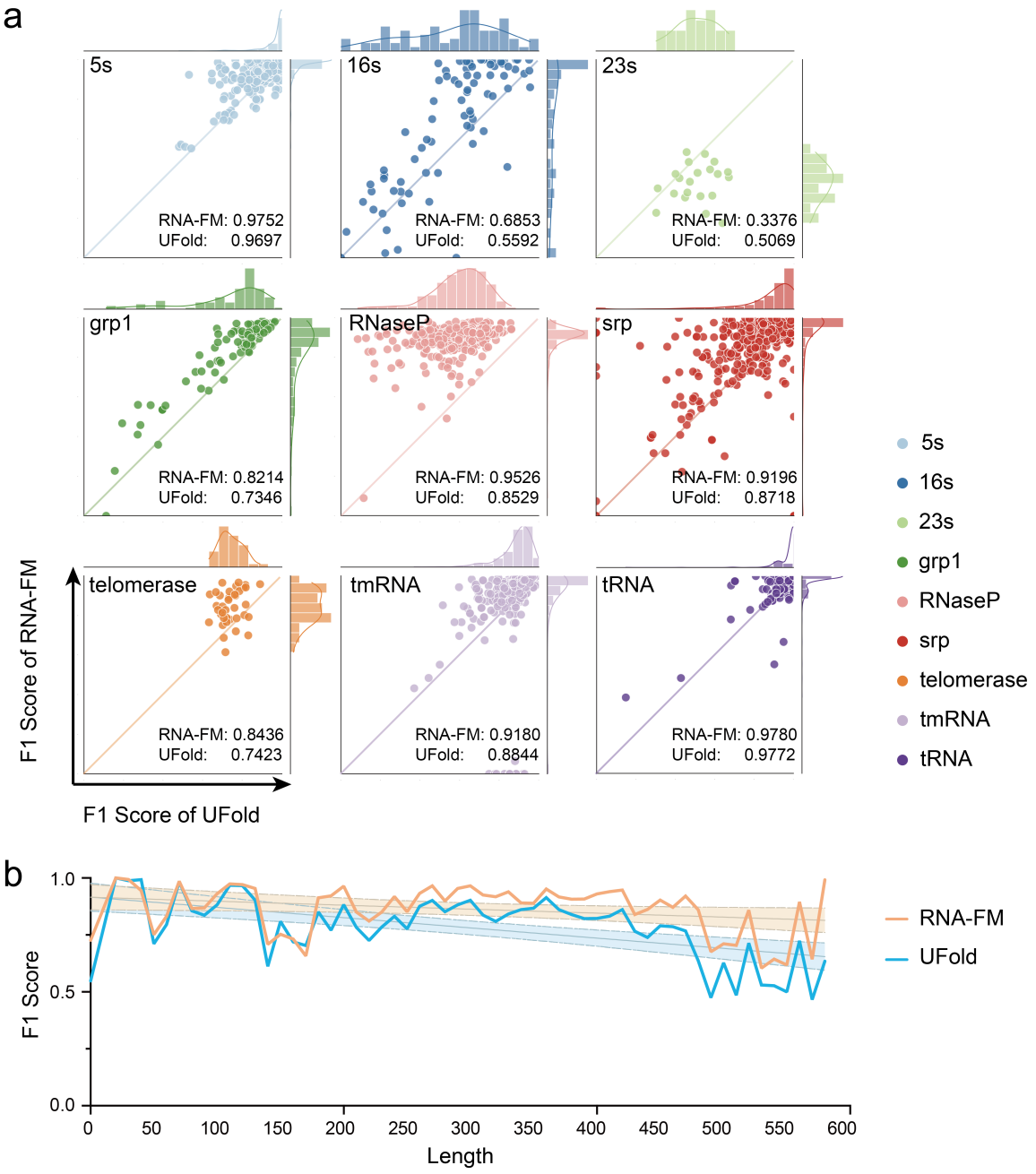
SARS-CoV-2 virus genome embedding extraction. We aim to extract representations for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) using RNA-FM. Considering the length of the whole genome (around 30K nucleotides) is far longer than the input limitation (1022 nucleotides) of RNA-FM, here, we apply a different feature extraction strategy. To be specific, we employ a fixed-length window of 1022nt on a non-overlap sub-section of the whole genome to extract the RNA-FM embeddings. Then, we aggregate embeddings by taking the average over them, and the final standard length vector is used as the RNA-FM embedding for the whole genome. We then apply trajectory inference with the genome-wise RNA-FM embeddings by VIA [42]. For each variant, we sample maximum 100 instances from the all sequences download from SARS-CoV-2 Data Hub. We set k in the k-nearest neighbors algorithm of VIA as 120 and the root as the Alpha variant.

Protein-RNA interaction. Protein-RNA interactions play important roles in a plenty of activities, such as cell-signaling, post-transcriptional regulations and protein synthesis [59]. Therefore, considering the importance

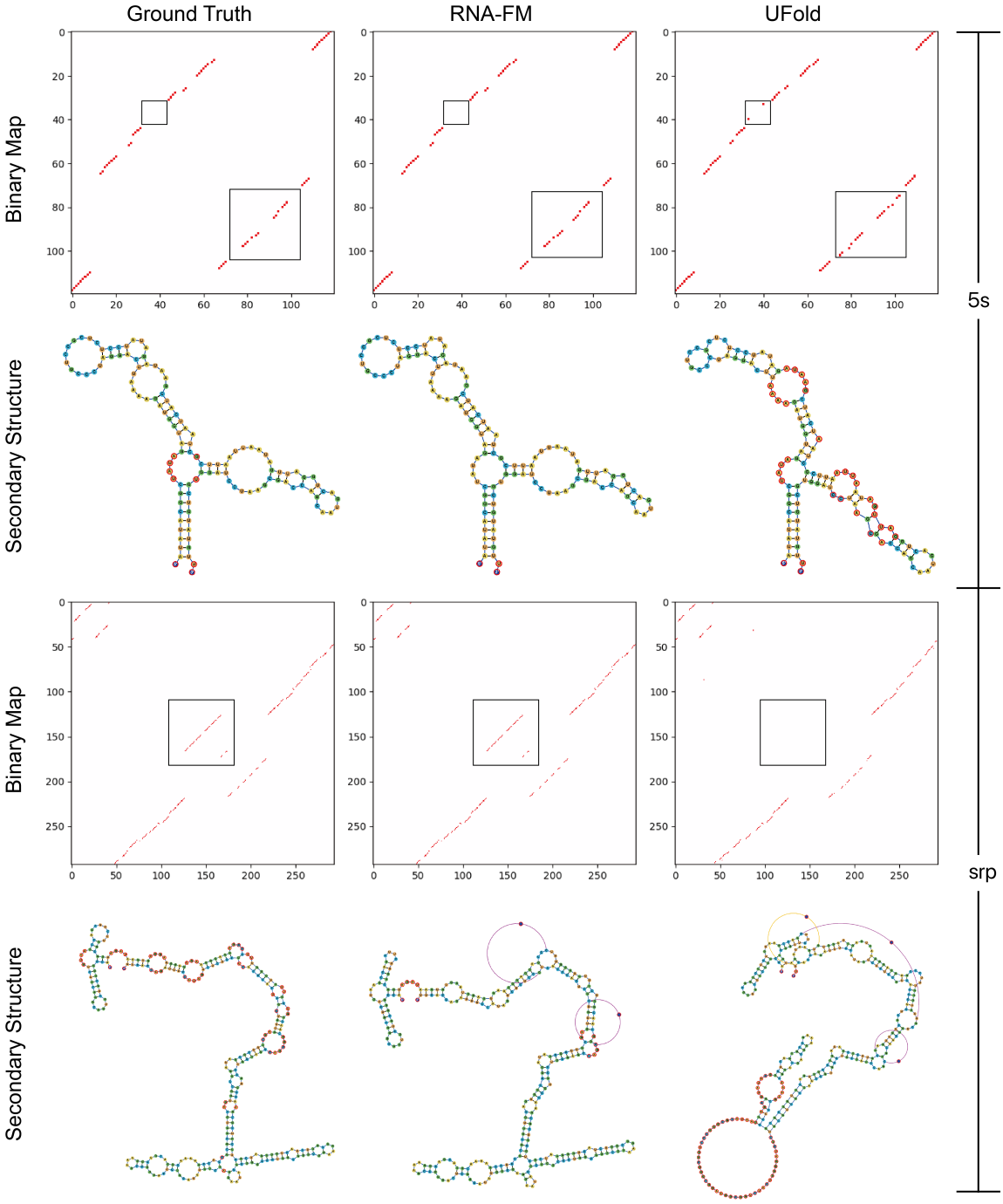
of protein-RNA interactions in RNA function, we evaluate how well RNA-FM mine ncRNA function information by predicting RNA binding proteins corresponding to RNAs. In this application, we apply Hela cell as the dataset for the task and devide it into 17 different protein sub-sets. Within each one, we predict whether the input RNA can bind with the protein. Our downstream prediction module perform similarly as the PrismNet [36] pipeline. Firstly, we reproduce PrismNet results, including two different types: One trained from raw sequences as baselines, and another one trained with *in vivo* secondary structures. We then replace the actual secondary structures in PrismNet with our generated RNA-FM embeddings for all sequences while keeping the whole downstream PrismNet architecture unchanged. In this way, we can test if using RNA-FM embeddings can accurately predict protein-RNA interactions as the real RNA secondary structures.

mRNA untranslated region's function. The 5' untranslated region is the region of a messenger RNA (mRNA). Although a 5'UTR is a part of a mRNA, which does not belong to ncRNAs, and RNA-FM is trained with ncRNAs, we attempt to test whether our pre-trained RNA-FM can handle these kinds of special non-coding sequences, and expect RNA-FM to finally aid with modeling the relationship between UTR and target protein expression. For the input UTR, we predict its corresponding mean ribosome load (MRL), which can reflect how a UTR regulates the target protein expression level. Specifically, we adopt the same pipeline and model as Paul et al. [37] developed in order to testify the effectiveness of our RNA-FM embeddings. Their model is well-designed by performing a grid search, whose best framework is constructed as three 1D convolutional layers with 120 filters and a ReLU activation for each layer. The third convolution layer will output one channel and L length features, which will be fed into two fully-connected layers with one output node as the final prediction. The original inputs of the model are simply the one-hot representation of raw sequences (4 dims). Therefore, to stay as close to their original architecture as possible, we replace the inputs with our embeddings. Finally, we obtain three models with different inputs, including pure sequence (Seq) in the form of one-hot encoding (4 dims), pure RNA-FM embedding (RNA-FM) of 640 dims, and the combination of these two (RNA-FM + Seq). Further, we apply a linear projection to reduce the embedding dimension from 640 to 4 for matching the one-hot embedding dimension.

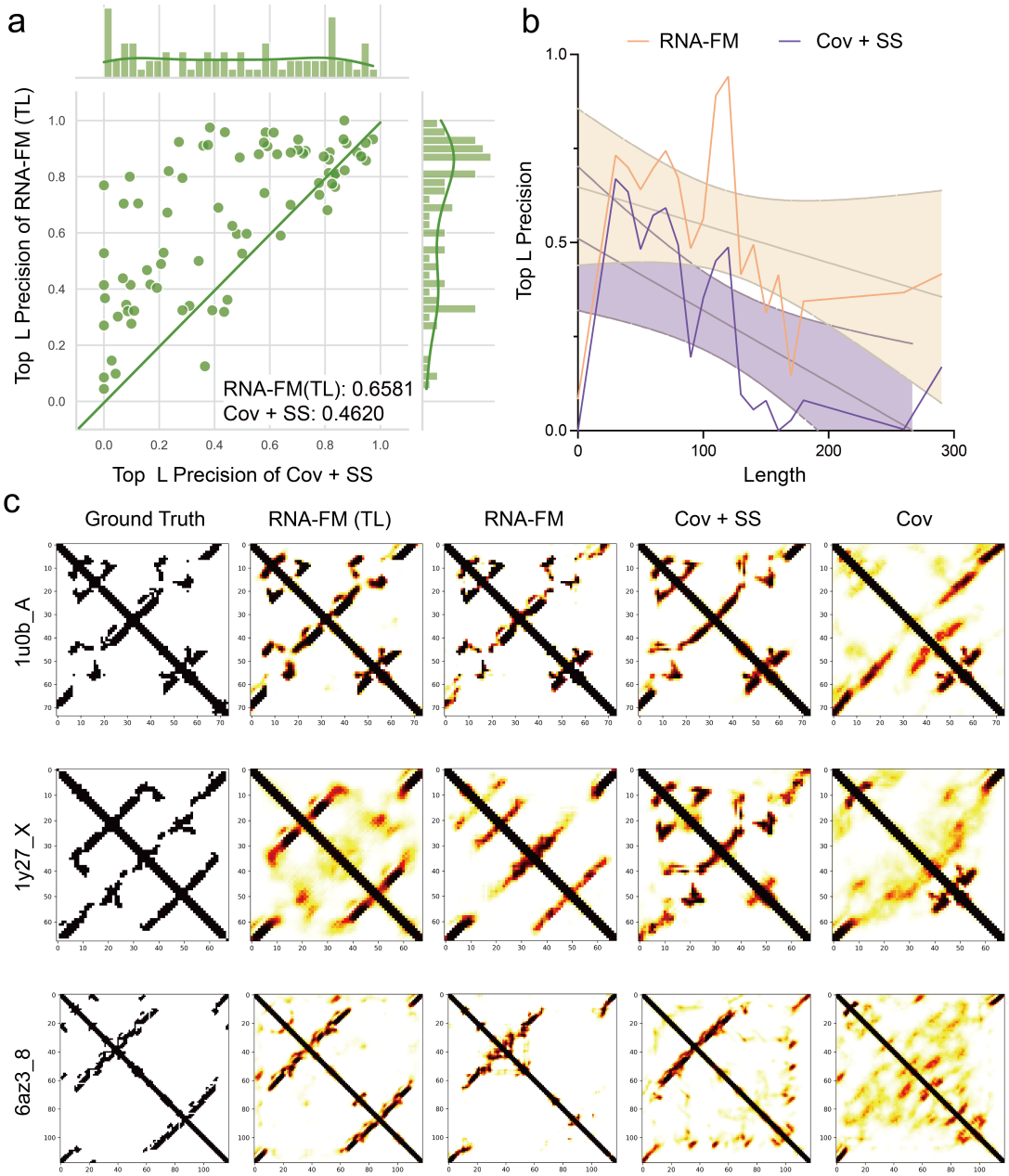
Appendix



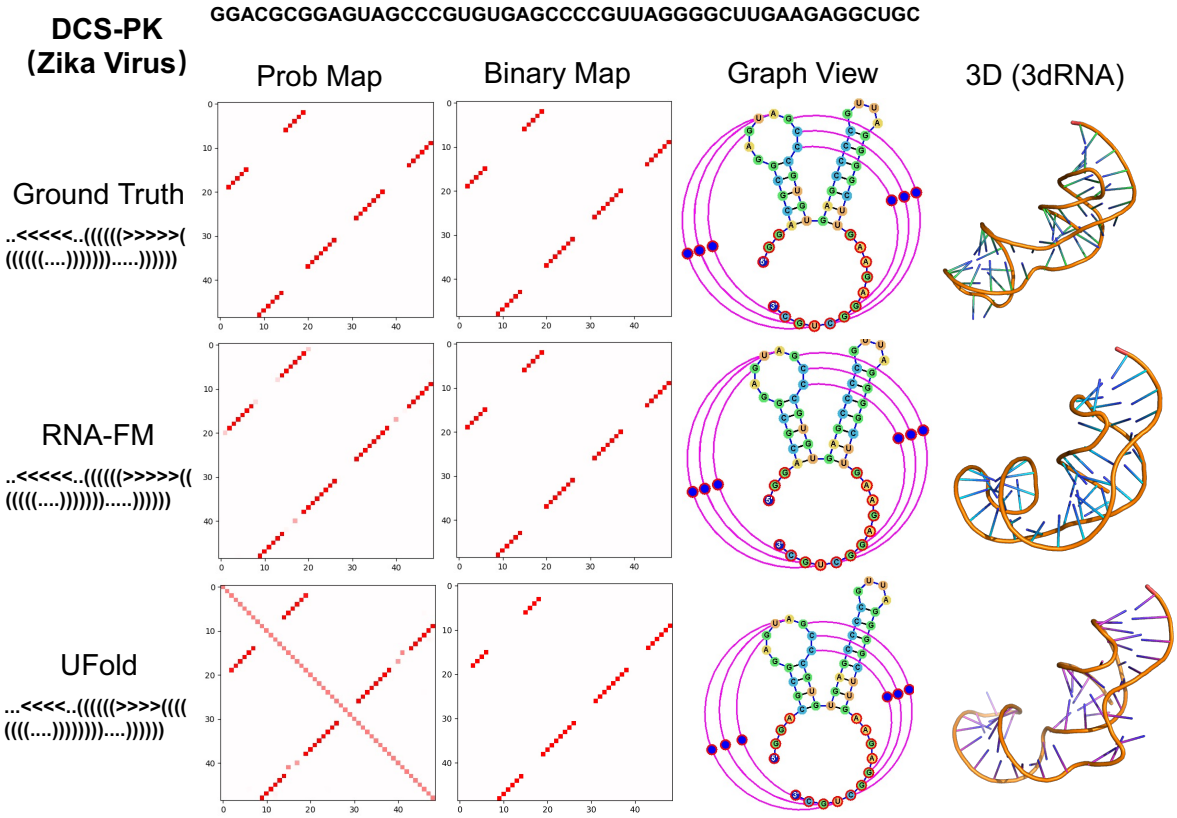
Appendix Figure 7: **Detailed performance comparison between RNA-FM and UFold on the ArchiveII dataset.** The experiment is cross-dataset validation of the trained model without re-training on ArchiveII. **a.** Scatter plots of F1 score comparison across 9 RNA types, with the performance of RNA-FM as the y-axis and that of UFold as the x-axis. Each point represents an RNA structure. Almost all the points are above the diagonal, which means RNA-FM beats UFold on nearly all the instances. **b.** F1 scores as a function of RNA sequence lengths. RNA-FM always outperforms UFold across all the lengths, especially when the length is over 150.



Appendix Figure 8: **Probability maps and graph view of secondary structure predictions of two randomly selected examples.** We compare the predictions from RNA-FM (second column) and UFold (third column) against the ground truth (first column). The probability maps from RNA-FM are more robust with less noise and closer to the ground truth compared to the ones from UFold. Regarding the visualization, RNA-FM also generates secondary structures more similar to the ground truth than UFold.



Appendix Figure 9: **RNA 3D closeness prediction performance on RNAcontact TE80 dataset.** ResNet is used to reproduce RNAcontact results for an equal comparison of different features. *Seq* means one-hot encoding of the sequence; *Cov* means MSA covariances; *SS* means secondary structure predicted by the PETfold based on MSA; *RNA-FM* means the feature from RNA-FM;+ means a combination of features by a channel-wise concatenation. **a.** Scatter plots of MCCs, with the performance of *RNA-FM* as the y-axis and the performance of *Cov+SS* (used by RNAcontact) as the x-axis. Each point represents an RNA structure. Almost all the points are above the diagonal, which means RNA-FM embeddings are better than the other features in nearly all the instances. **b.** F1 scores as a function of RNA sequence lengths. *RNA-FM* outperforms *Cov+SS* all the time. **c.** Probability maps of three randomly selected examples. With the standalone RNA-FM embedding (*RNA-FM*) as the input, the downstream model has already generated visualizations much closer to the ground truth than other features. Furthermore, by re-training the ResNet pre-trained on the other tasks using transfer learning, we can achieve the performance far better than the other methods.



Appendix Figure 10: **3D reconstruction of RNA.** The probability maps and binary maps are generated by different predictors. The graph views are obtained by jViz.Rna 4.0 [50]. The 3D structures are modelled by 3dRNA. **a.** An instance from PDB (5m73-1-A). **b.** The DCS-PK of Zika Virus.