ProtRNA: A Protein-derived RNA Language Model by Cross-Modality Transfer Learning

Ruoxi Zhang[1], Ben Ma[1,2], Gang Xu[2,3*], & Jianpeng Ma[2,3*]

[1]*Fudan University*
Shanghai, 200433, China

[2]*Multiscale Research Institute of Complex Systems*
Fudan University
Shanghai, 200433, China

[3]*Shanghai AI Laboratory*
Shanghai, 200030, China

August 28, 2024

[*]Correspondence: xugang@fudan.edu.cn, jpma@fudan.edu.cn

Abstract

Protein language models (PLM), such as the highly successful ESM-2, have proven to be particularly effective. However, language models designed for RNA continue to face challenges. A key question is: can the information derived from PLMs be harnessed and transferred to RNA? To investigate this, a model termed ProtRNA has been developed by cross-modality transfer learning strategy for addressing the challenges posed by RNA's limited and less conserved sequences. By leveraging the evolutionary and physicochemical information encoded in protein sequences, the ESM-2 model is adapted to processing "low-resource" RNA sequence data. The results show comparable or even superior performance in various RNA downstream tasks, with only 1/8 the trainable parameters and 1/6 the training data employed by other baseline RNA language models. This approach highlights the potential of cross-modality transfer learning in biological language models.

Introduction

As the Transformer architecture continues to dominate state-of-the-art (SOTA) performance in natural language processing (NLP) tasks, its applications to biological sequences have gained increasing attention. By treating the building blocks of biological sequences as “tokens”, researchers have developed large Biological Language Models (BLM) pretrained on DNA, RNA or protein sequence data, which are being deployed for a wide range of structural and functional prediction tasks.

BERT-style [1] BLMs, which are pretrained by Masked Language Modeling (MLM) and are able to produce bi-directional representations, have recently gained unignorable advancements. Notably, BERT-style protein language models, like ESM-2, have excelled in structure prediction [2] and various other downstream tasks. These models are thought to internalize evolutionary patterns by pretraining on large protein sequence databases [3]. The bi-directional self-attention mechanism in these models is suggested to enable capture of residue-residue contacts, storing this information in attention heads and representations [4].

Unlike protein language models, such as the aforementioned SOTA ESM-2, few general-purpose RNA language models show robust performance. RNA-FM [5] is one of the first RNA foundation language models trained on noncoding RNA (ncRNAs). It consists of 12 standard encoder transformer blocks, with 640 embedding dimensions and is trained on 23 million sequences from the RNAcentral database [6]. RiNALMo [7] is a larger model with 33 transformer blocks, 1280 embedding dimensions and is trained on a curated 36 million ncRNA dataset, with some architectural modifications that incorporate rotary positional embedding (RoPE) and SwiGLU activation function.

Nonetheless, the effectiveness of BERT language modeling in capturing evolutional information in RNA has been questioned [8]. RNA language modeling is suggested to be intrinsically more challenging than that of protein, due to the limited four-letter alphabet and the lower RNA sequences [9]. This results in sparse and noisy data, making RNA sequences more

difficult to model effectively. The low quality of data in RNA language thus renders it a "low-resource" language in NLP terminology, in contrast to the 'resource-rich' protein language.

Following the line of cross-lingual transfer learning studies in the field of NLP [10], we explored a cross-modality transfer learning pretraining strategy that leverages the SOTA ESM-2 model trained on the 'resource-rich' language protein sequences and adapts it to 'low-resource' ncRNA data (**Figure 1.a**). Our approach leverages the evolutional and physicochemical information encoded in the pretrained protein ESM-2 model to facilitate RNA language modeling. The resulting ProtRNA model, derived from this transfer learning pretraining strategy, exhibits performance that is comparable with or even superior to that of existing foundational RNA language models across various RNA downstream tasks (**Figure 1.b-d**). Remarkably, this was achieved with merely 1/8 trainable parameters and 1/6 training data employed by other comparable models, underscoring the efficiency and effectiveness of our approach.

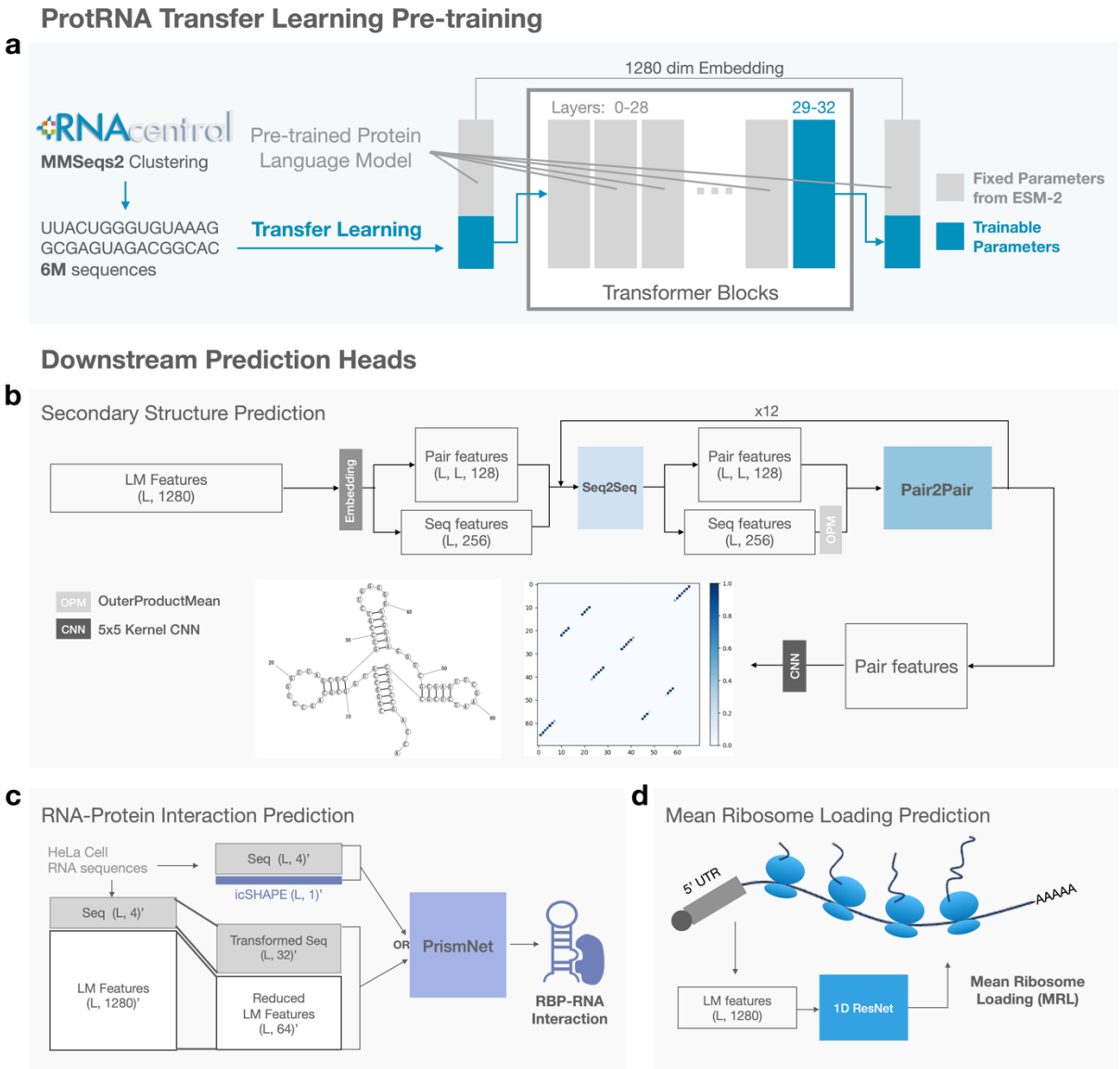

Figure 1. Overview of the ProtRNA transfer-learning pretraining strategy for RNA language model and downstream task architectures. a, The ProtRNA model utilizes initial weights from the SOTA protein language model ESM-2. During the cross-modality transfer learning phase, the model is pretrained on 6 million non-coding RNA sequences using the MLM task. It freezes the first 29 transformer layers, updating only RNA vocabulary embeddings and the last four layers (30-33). **b-d,** After pretraining for six epochs on RNA data, the ProtRNA model is evaluated on three downstream RNA tasks. For these tasks, task-specific heads receive sequence representations from the pretrained language model, and only the heads are further trained on respective training sets.

Results

Cross-Modality Transfer Learning Strategy for ProtRNA Pretraining

Parameter and Data Efficient Pretraining. Utilizing the ESM-2 model with 1280 embedding dimensions and 33 transformer layers, we froze the first 29 layers and trained only the final 4 layers and the embeddings for RNA tokens. Training was performed on 6 million representative RNA sequences clustered by MMSeqs2 [11] from the RNAcentral database. As could be seen from **Table 1**, compared with other BERT-style RNA language models, both the trainable parameters and the effective training data utilized in our pretraining process are considerably fewer. We will show in the following section that even with this amount of reduction in computational resource, the performance of our model in various downstream task is comparable with or even superior to other models. The efficiency of our transfer learning method is also demonstrated in the ablation study on a smaller dataset, shown in **Figure S1**, where it is clear that the training loss and validation accuracy improve significantly faster than training from scratch with the same number of trainable parameters and randomly initialized weights.

Table 1. Comparison of Training Process and Model Architecture between ProtRNA and other foundational RNA language models. The dash ('-') indicates the information is not disclosed in the publication. The Uni-RNA model, not released publicly, is excluded from subsequent downstream task testing.

	ProtRNA	RiNALMo	Uni-RNA	RNA-FM
Trainable Params	80M	650M	400M	100M
Training Data	6M	36M	500M	23M
Total Params	650M	650M	400M	60M
Total Layers	33	33	24	12
Max Length	512	1024	-	1024
Embedding Dim.	1280	1280	1280	640

Distinguishable Clustering of RNA Families and Protein-RNA Segregation in Representation Space. In **Figure 2**, we present a t-SNE-based 2D visualization of the RNA and protein sequence representations generated by our final ProtRNA model. The RNA sequences are from the ArchiveII dataset [12], and protein sequences are from UniProt [13]. The t-SNE algorithm effectively captures the local and global structure of the high-dimensional sequence data in a lower-dimensional space [14]. The visualization reveals that RNA sequences from different families form distinct clusters. Additionally, the protein representations are clearly segregated from the RNA representations, residing in the upper left corner of the plot. This outcome highlights the model's ability to discriminate between RNA families, while also effectively distinguishing protein from RNA representations.

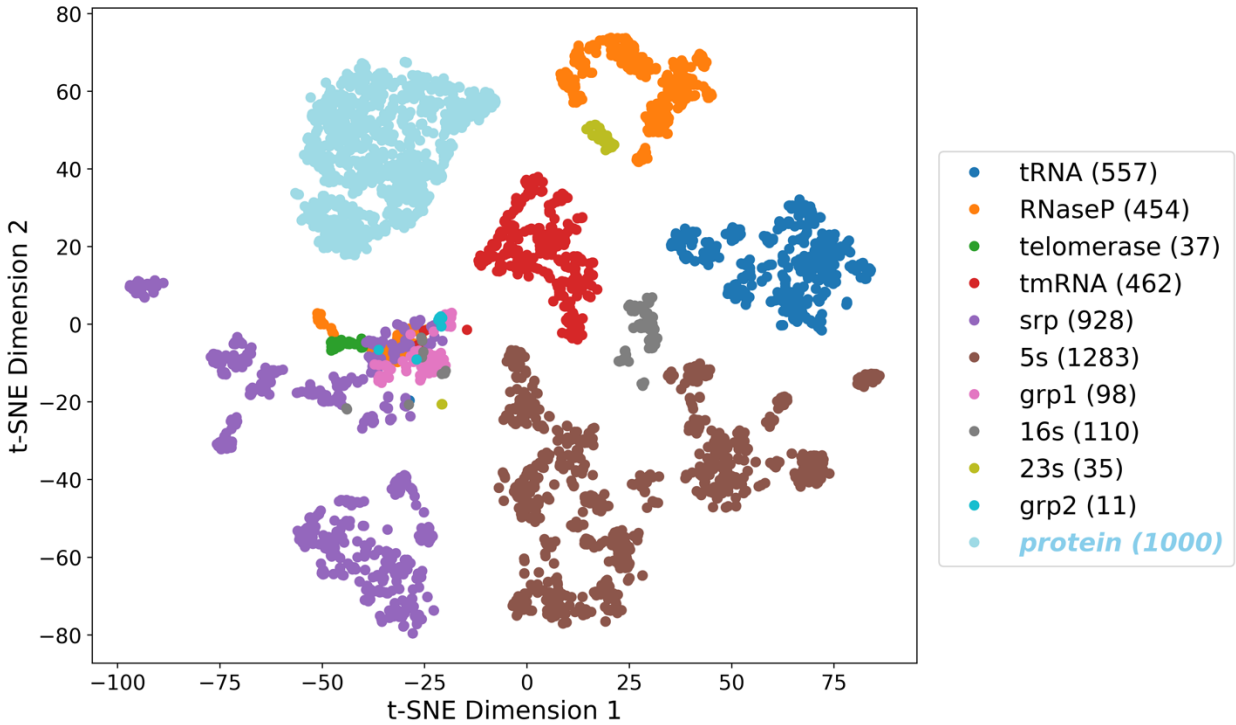

Figure 2. **t-SNE visualization of ProtRNA Representations.** RNAs from different families are colored differently. The protein cluster (denoted in skyblue) is distinctly separated from the RNA families at the upper left corner.

Distinctive Attention Pattern in the ProtRNA Model Where a Specialized Structure Head Emerges. Though interpretability of BERT models in NLP, particularly how they encode dependency relations through self-attention, has been extensively studied [15] [16] [17] [18],

research on the inner workings of BERT models in the context of biological sequences is relatively sparse. The multi-head self-attention mechanism is thought to be crucial, as it enables the model to capture pairwise dependencies within a sequence, especially long-range interactions, which may correspond to the spatial contacts that occur when a sequence folds into a stable three-dimensional structure [19]. Hereby we provide an attention analysis of our ProtRNA model and two other baseline RNA BERT models, with regard to two structural properties of RNA, applying the method devised by [4]. See **Methods** for details of the metric and the datasets used.

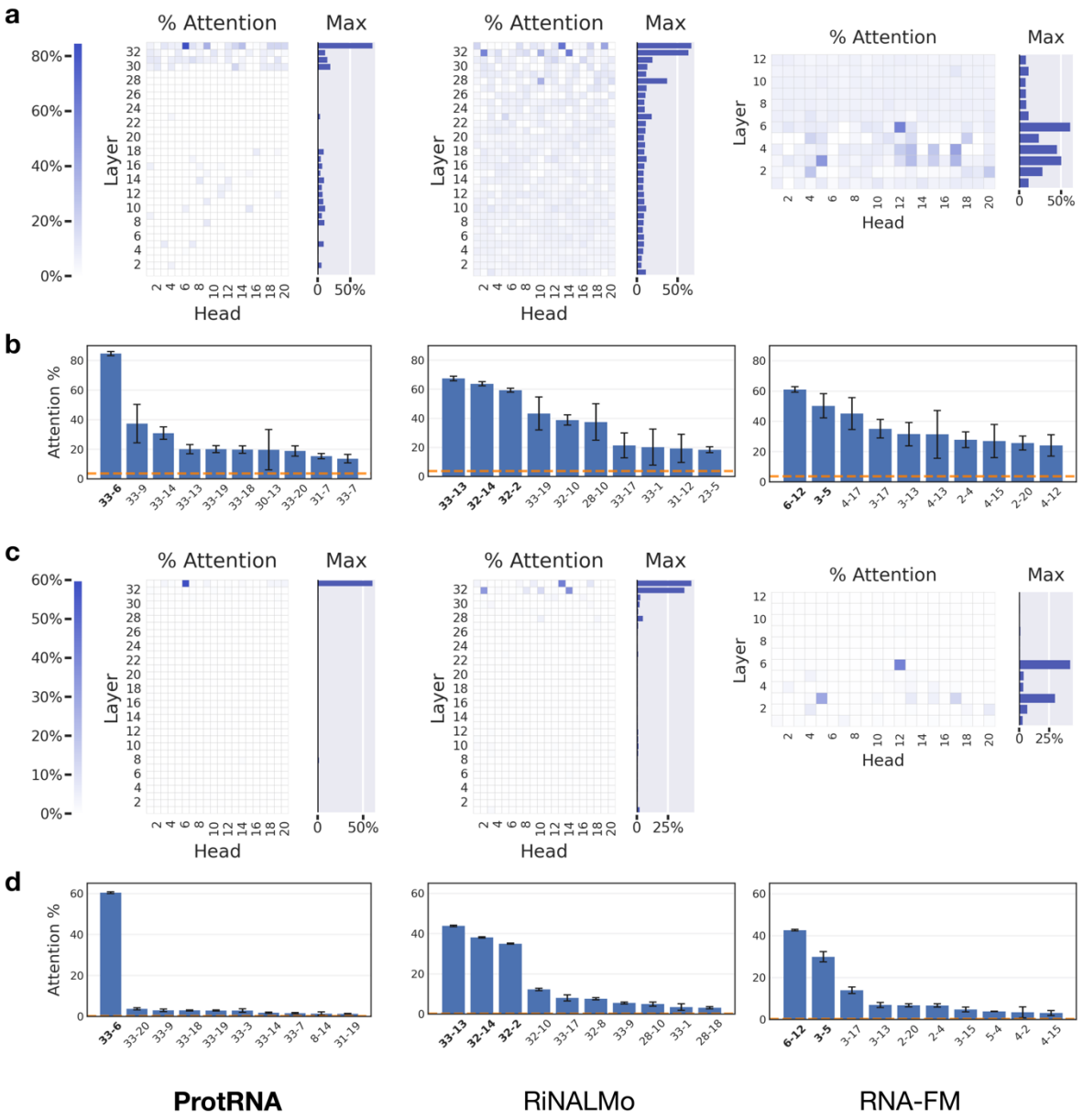

Figure 3. Attention analysis of ProtRNA against two baseline RNA language models. a-b, the results on contact map property. **c-d,** the results on secondary structure property. Attention% score for each head measures the proportion of its high attention weights that align with pairs in close contact among sequences from the RNA contact map dataset, or which form hydrogen-bond across sequences from RNA secondary structure dataset. The bar plots **b** and **d** under the heatmaps show top 10 heads in the model with highest scores, where the orange dashed lines represent background frequency of the property under examination, and the black lines denote confidence interval of the estimated scores under significance 0.05 with Bonferroni correction for multiple hypotheses.

As shown in the plot, ProtRNA exhibits a distinctive attention head pattern compared with the two other general RNA language models. Notably, ProtRNA features a highly specialized attention head, head 33-6, for both contact map and secondary structure properties. This head achieves a remarkable score of 84.6% for the contact map property, indicating that 84.6% of the high attention weights in this head correspond to pairs in contact. In contrast, RiNALMo and RNA-FM, while also employing shared structural heads for the contact map and secondary structure properties, depend on multiple heads with relatively lower scores. Specifically, RiNALMo mainly utilizes three such heads and RNA-FM employs two. This difference might stem from our constrained transfer learning strategy, where the ProtRNA model is required to compress information efficiently within just four unfrozen layers during pretraining on RNA sequence data.

Evaluation of ProtRNA Representations on Downstream Tasks

Secondary Structure Prediction Using an Adapted Protein Structure Prediction Head. The secondary structure of an RNA molecule is the set of hydrogen-bond nucleotide base pairs formed within its primary sequence, including the Watson-Crick pairs A-U, G-C, and the Wobble pair G-U. Such base-pairing interactions induce secondary substructure motifs like stackings, loops, and bulges. Among the four levels of hierarchy in RNA structure, the secondary structure reveals basic information about the function of the molecule, characterizes the homology family it belongs to [20] [21], and facilitates construct design for tertiary structure determination [12].

Among all the deep learning models trained via direct supervised learning on secondary structure prediction, the following three models showed comparatively good results consistent across reports from different sources. SPOT-RNA deploys an ensemble of two-dimensional deep neural networks combining Residual Networks (ResNets) and Bidirectional Long Short-Term Memory cells (2D-BLSTMs) to output a 2D base-pair contact probability matrix [22]. The current SOTA DNN model UFold implements a U-Net architecture widely used in image segmentation tasks, to directly obtain a 2D contact map as SPOT-RNA, and apply post-process techniques to refine the results [23]. MXFold2 integrates deep learning with thermodynamic methods by folding in Zuker's style to maximize the sum of Turner's nearest neighbor free-

energy parameters and folding scores computed from a deep neural network with convolutional layers and BLSTMs [24].

Another DNN-based methodology leverages transformer language models for sequence representation, similar to their application in protein structure prediction. A common practice in previous language models for secondary structure prediction involves training a relatively simple prediction head, such as the two-layer ResNet bottleneck used in RiNALMo. This approach exploits the self-attention mechanism, which constructs pairwise interactions between all positions in the sequence. The inductive bias inherent in this architecture effectively represents base pair interactions, facilitating accurate secondary structure predictions.

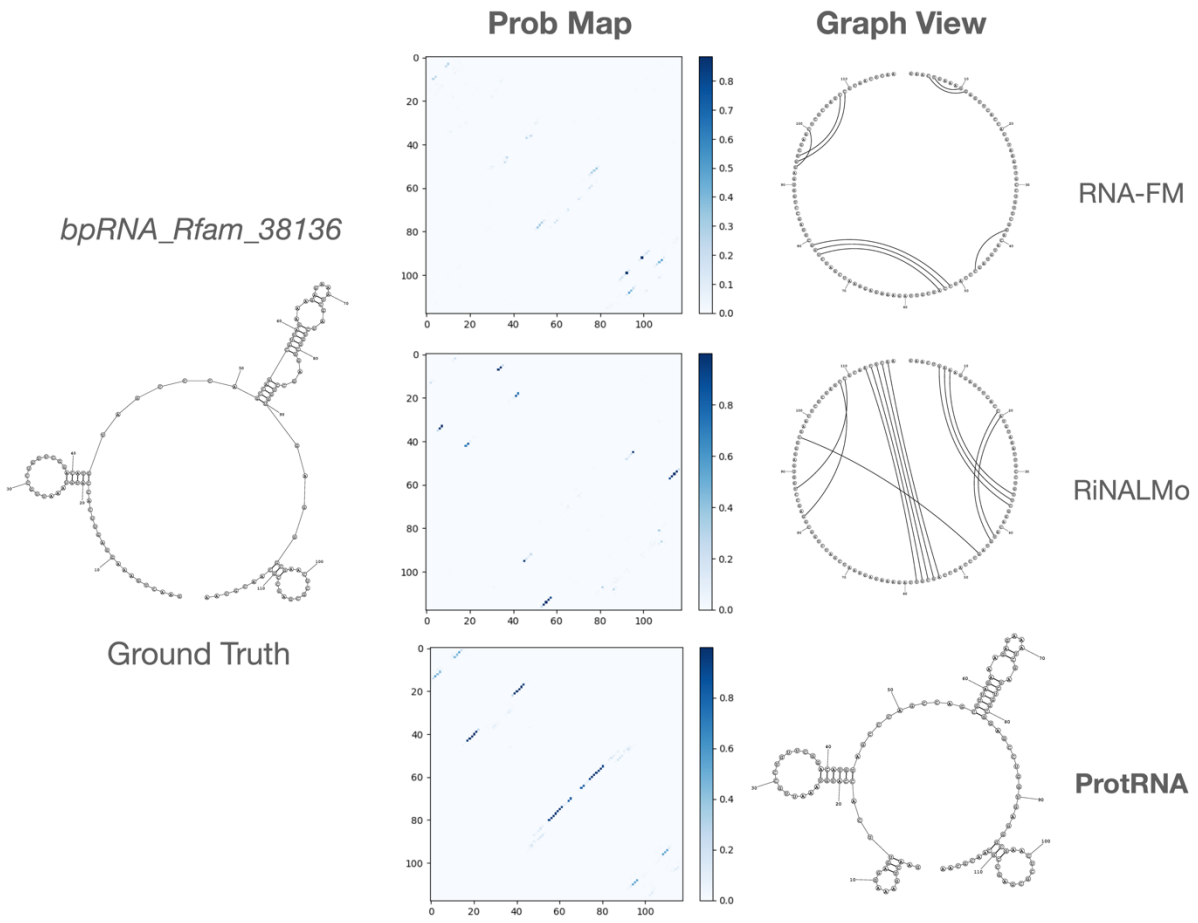

Figure 4. The secondary structure probability maps and binary pairing in graph view from the prediction of the three RNA language models on a sequence from bpRNA dataset.

Here we propose a new prediction head (**Figure 1.b**), which is an adapted AlphaFold2-style RotaFormer module from OPUS-Rota5 [25], originally used for highly accurate protein contact prediction. This advanced architecture has shown to be more effective in extracting the embedded structural information in the language model representations.

We compared the performance of domain-specific models and pretrained language models with two different prediction heads on the bpRNA-1m dataset under our evaluation pipeline. To compare the effectiveness of the RNA language model representations in this task, we utilized the raw sequence representations from the pretrained neural networks, with no fine-tune or training dropout as stated before. The two heads for the language models are the two-layer ResNet head from RiNALMo along with its post process techniques, and the adapted RotaFormer head with no post process. The language model heads are all properly trained and tested using the train-val-test split dataset proposed in SPOT-RNA. For the domain-specific models, we tested on their released model, with training set as close to TR0 as possible for fair comparison. The post process techniques are left to their own original architectural design.

Table 2. RNA secondary structure prediction performance of domain-specific models and language models with two types of heads evaluated on TS0. MCC refers to the Mathews Correlation Coefficient, and F1-shifted refers to the F1 score obtained after allowing flexible pairing.

	Precision	**Recall**	**F1**	**MCC**	**F1-shifted**
Domain-Specific DNN					
SPOT-RNA	0.5784	0.6522	0.5965	0.6039	0.6394
MXfold2	0.5224	0.6214	0.5537	0.5602	0.6018
UFold	0.6106	0.6769	0.6306	0.6356	0.6886
Language Model					
RiNALMo	0.5997	0.5936	0.5864	0.5898	0.6354
RiNALMo-our-head	0.6263	0.6266	0.6163	0.6199	0.6684
RNA-FM-our-head	0.6691	0.6067	0.6225	0.6285	0.6718
ProtRNA	0.6472	0.6515	0.6380	0.6422	0.6940

From **Table 2**, it is evident that our model outperforms others on the intra-family secondary structure prediction task. This highlights its accuracy in predicting RNA secondary structures in the intra-family context even with significantly reduced trainable parameters and required training data of the pretrained language model. Additionally, our AF2-style RotaFormer head surpasses the RiNALMo head in leveraging language model features. The RotaFormer head's advanced design enables better utilization of sequence information, enhancing prediction accuracy. In contrast, the RiNALMo head's simpler two-layer ResNet bottleneck is less effective in capturing the complex relationships in the data. This demonstrates the superiority of our approach in integrating and interpreting sequence features. **Figure 4** demonstrates an example where our ProtRNA with RotaFormer head predicts RNA secondary structure closer to ground truth than the other two language models. The graph view in **Figure 4** is generated by RNAStructure [26].

Protein-RNA Interaction. The sequence of a biomolecule dictates its cellular function metaphorically to how code specifies a program, where the sequence acts as the code, which is 'compiled' into a unique structure through folding, allowing the biomolecule to interact with other molecules and execute its function within the cell [27]. Thereby it is of interest how the sequence representations, yielded by compiling the sequence code to an extent through our transfer-learned language model and are showed above to contain structure information, could further facilitate tasks involving RNA interactions with other molecules in the cell.

Predicting RNA-binding protein (RBP) binding sites involves identifying specific regions on RNA molecules where RBPs interact. These interactions are crucial for various cellular processes, including RNA splicing, transport, stability, and translation [28] [29] . The task is complex due to the dynamic and context-specific nature of RNA structures and RBP interactions. Previous domain-specific DNN method, PrismNet [30], addresses the limitations of traditional methods by integrating deep learning with in vivo RNA structural data obtained using in vivo click Selective 2′-Hydroxyl Acylation and Profiling Experiment (icSHAPE), which allows the model to capture dynamic and condition-specific interactions between RBPs and RNA.

Here we reproduced PrismNet but substituted the icSHAPE experimental RNA structural data with RNA sequence representations obtained from our ProtRNA model and other benchmarking pretrained RNA language models (see **Figure 1.c**). Using the HeLa cell dataset, we aim to test if our ProtRNA sequence representations could effectively facilitate the RBP prediction task.

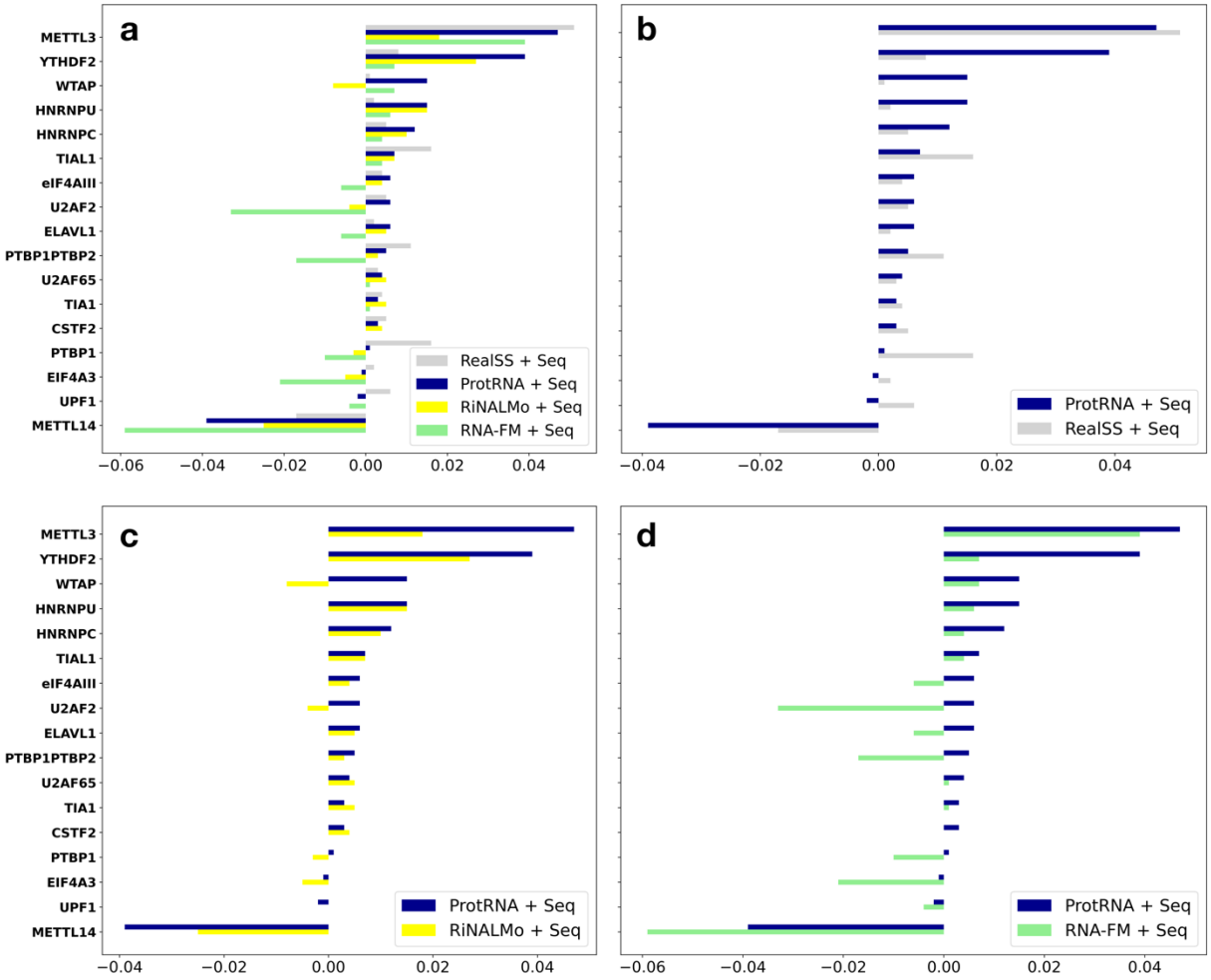

Figure 5. Protein-RNA interaction performance (AUROC) of RNA language model features on 17 RBPs in Hela Cell. "Seq" stands for one-hot representations of the RNA sequence. "RealSS" refers to the 1-dimensional experimental icSHAPE data of the RNA secondary structure. "+" means the features are concatenated as inputs to the PrismNet model. The y-axes are the 17 RBPs, and the x-axes are the area under the receiver operating characteristic (AUROC) score of each model. The AUROC scores are shifted to compare with the model performance with inputs of "Seq" only.

As shown in Figure 4, compared with the baseline seq-only model, ProtRNA + Seq consistently demonstrates significant positive impacts and only minor negative impacts in most cases. This performance matches or surpasses that of the RealSS + Seq model (see **Figure 5.b**), which utilizes icSHAPE RNA structural data. ProtRNA + Seq also outperforms the other two models (see **Figure 5.c-d**): RNA-FM exhibits notable negative impacts for many proteins, and RiNALMo largely fails to achieve the improvements seen with ProtRNA or RealSS. The performance on the protein-RNA interaction task suggests that, despite using less training data on RNA sequences and fewer trainable parameters, our pretraining strategy enhances the model's ability to handle the complex interaction task between proteins and RNA. This may be attributed to the pretraining process that incorporates biomolecules of different modalities.

Mean Ribosome Loading. Mean Ribosome Loading (MRL) refers to the average number of ribosomes bound to a single mRNA molecule during translation. It measures translation efficiency, indicating how actively a particular mRNA is being translated into protein. The 5' untranslated region (5' UTR) of mRNA regulates protein translation efficiency, as it contains specific sequences and structural motifs crucial for the binding of ribosomes and translation initiation factors [31] [32]. Accurate prediction of MRL from the 5'UTR of an mRNA sequence would significantly enhance drug design in mRNA therapeutics.

Predicting MRL from 5'UTR is a regression task that involves creating models capable of capturing the regulatory features within the 5'UTR and predicting their impact on ribosome loading. Advanced machine learning techniques, particularly deep learning models, are employed to learn these complex relationships from large datasets of 5′ UTR sequences and their corresponding MRL values [33] [34].

In this task, we aim to test our model's generalizability to the structural information contained in the 5'UTR of mRNA sequences, which is a noncoding region that serves a function role. The training dataset and the evaluation datasets, Human7600 and Random7600 are provided by [35]. We utilized the same prediction head architecture as in Uni-RNA and RiNALMo, featuring a simple 1D ResNet (see **Figure 1.d**). For all three RNA language models

under comparison, the sequence representations are from pretrained models, and only the prediction head is trained. During training, R^2 is used as the evaluation metric, as in RiNALMo.

Table 3. MRL prediction performance evaluation on 5'UTR sequences from Random7600 and Human 7600 dataset.

Model	**Random7600**		**Human7600**	
	R^2	MAE	R^2	MAE
RNA-FM	0.5843	0.6488	0.6472	0.5052
RiNALMo	0.5828	0.6521	0.6480	0.5091
ProtRNA	0.6060	0.6243	0.6738	0.4862

As seen from **Table 3**, our ProtRNA representations of the 5'UTR achieve higher R^2 and lower Mean Absolute Error (MAE) in the regression task. This indicates the better generalization ability of ProtRNA model to RNA sequences out of distribution of the pretraining data, compared with the other two RNA language models that are also trained solely on ncRNA data. These results underscore the generalizability of our ProtRNA representations.

Discussion

In this paper, we introduced a cross-modality transfer learning strategy for RNA language model pretraining, leveraging model parameters from ESM-2, which is extensively trained on the protein language, a 'resource-rich' modality. Our ProtRNA model, built on this transfer learning approach, achieves comparable or even superior performance in several RNA downstream tasks when compared to other RNA foundation models and domain-specific deep learning models. Such performance highlights that the effectiveness of transferring information from a pretrained protein language model to RNA sequences.

Specifically, a new secondary structure prediction architecture is proposed, shown in **Figure 1.b**, which combines language model representations with an advanced AF2-style RotaFormer module. This architecture demonstrates superior performance (**Table 2**), adding up

to the DNN-based toolbox in RNA secondary structure prediction. The performance of ProtRNA representations in the protein-RNA interaction task (**Figure 4**), highlights the effectiveness of the encoded information, resulted from evolutional transfer learning pretraining, to facilitate task involving the interaction between biomolecules of different modality. Finally, ProtRNA generalizes to produce mRNA 5'UTR representations that achieves better MRL prediction accuracy (**Table 3**).

Additionally, since the ProtRNA model achieves competitive performance in RNA-specific tasks while retaining most of the transformer layer parameters from the protein model ESM-2, it may serve as an intriguing subject for future studies on the interpretability of deep learning models within a biological context. Two preliminary analyses were conducted: the t-SNE plot demonstrates the ProtRNA representation space with clearly segregated protein and RNA sequences; the attention analysis reveals a distinctive attention head pattern of the ProtRNA model related to RNA structural properties. While further analysis to understand the mechanisms behind this performance is beyond the scope of this work, it is a promising avenue for future research that could yield insights into the interplay between different biological modalities and inform the development of improved models.

For our proposed cross-modality transfer learning strategy, the current model only utilizes data from the target modality, in our case, RNA. To optimize the model performance on the target modality, other established methods in cross-lingual transfer learning may also be adopted, such as the representation alignment or joint training objectives between two languages [36] [37]. However, it is crucial to note that the mapping relationship between protein and ncRNA sequences differs significantly from that in natural languages. Therefore, adapting these NLP methods requires careful consideration. For instance, when constructing training batches with sequence pairs of ncRNAs and proteins to analogously match their "meaning", one might consider incorporating structural or functional motifs. In contrast, for mRNA, the mapping between proteins and mRNA follows a straightforward codon rule, which could be captured easily by a well-trained neural network. Thus, alignment training between these two modalities may not add significant semantic value. Taken the above considerations into account, more

delicate cross-modality training strategies suitable for protein and RNA, or even other modalities, could be explored in the future.

Methods

ProtRNA Pretrained Language Model

Pretraining Data. The original dataset, collected from the latest RNAcentral release when the experiments are conducted, contains 38 million RNA sequences. To reduce sequence redundancy, we applied MMSeqs2 algorithm with parameters set to form clusters with 90% sequence identity and representative sequences covering at least 95% of the cluster members. This process reduced the dataset to 6 million representative sequences, resulting in an 84% reduced sparse dataset. This reduction is comparable to the UniRef50 dataset used for ESM-2 pretraining, which achieved a 79% reduction from the original UniRef100 dataset.

Tokenizer. Similar to other general ncRNA language models, RNA sequences in ProtRNA are tokenized at single-nucleotide resolution, treating each nucleotide base as a separate token. The vocabulary expands upon the original ESM-2 amino acid tokens by including nucleotide tokens ("a", "c", "u", "g", and "x"). Lowercasing distinguishes nucleotide tokens from uppercase amino acid tokens, and the "x" token represents other non-common modified or unknown nucleotides. Special tokens like <cls>, <pad>, <eos>, and <mask> utilize the same encoding and embeddings as in ESM-2 to leverage learned embeddings.

Transfer Learning Pretraining Method. For transfer learning pretraining, we leveraged fixed parameters from the 650M parameters ESM-2 model, with 33 transformer blocks and 1280-embedding dimension, that was sufficiently pretrained on protein data UniRef50/D. To fully utilize the evolutional information stored in its attention head weights, we applied minimum modifications to the original model architecture. Namely, we only extended the embedding layer and the bias layer to accommodate the size of added RNA tokens. During training, most of the transformer blocks in the protein model ESM2, i.e., 29 among 33 (≈ 88%) of the blocks, were fixed.

We allow a context window of 512. Our model is pretrained on 4 V100 GPUs of 32GB memory for 6 epochs. During 5000 warm-up steps, we gradually increase the learning rate to 1e-4 and use Adam optimizer, then Nadam optimizer is applied with the constant learning rate 1e-4. The pretraining task for our architecture is as suggested in RoBERTa [38], where masked language modeling is applied with 15% tokens set to be targets and replaced with 80% the time a <mask> token, 10% a random token from the vocabulary, and 10% the original token for robust learning. The loss function is the classic cross-entropy loss for multi-class prediction. And the weight updates are based on the average loss per batch.

Attention Analysis

Metric. The metric $p_\alpha(f)$ measures the alignment of attention weights with underlying properties of token pairs, in our case, contact map or secondary structure base pairing. Indicator function $f(i,j)$ returns 1 if the property is present in token pair (i,j), 0 otherwise. θ is the threshold for high attention weights. For each attention head, the score is aggregated for all sequences x over the dataset X. In our analysis, the threshold θ is set to be 0.3, following the practice in [4].

$$p_\alpha(f) = \frac{\sum_{x \in X} \sum_{i=1}^{|x|} \sum_{j=1}^{|x|} f(i,j) \cdot \mathbb{I}_{\alpha_{i,j} > \theta}}{\sum_{x \in X} \sum_{i=1}^{|x|} \sum_{j=1}^{|x|} \mathbb{I}_{\alpha_{i,j} > \theta}}$$

Dataset. The contact map dataset is the whole dataset as provided in the RNA benchmark BEACON [39], while the secondary structure dataset is composed of 5000 sequences randomly chosen from the training set TR0 of the bpRNA dataset.

Baseline RNA Language Models for Downstream Evaluation

In this paper, we chose RNA-FM and RiNALMo as baseline RNA language models, which are open-resource models with reported competent performance. When evaluating the model performance on downstream tasks, we have restrained from fine-tuning the base model or employing typical training techniques such as dropout. These restraints ensure a fair comparison among foundational RNA language models, thereby providing a clearer indication of the quality of the extracted features.

Secondary Structure Prediction

Dataset. bpRNA-1m is a benchmark in RNA secondary structure prediction for deep learning methods, with the training, validation and test split proposed in SPOT-RNA [22]. It contains over one million structures collected from seven sources, including consensus structures derived from comparative sequence analysis and also X-ray crystallography and NMR techniques from Protein Data Bank (PDB).

RotaFormer Head. As illustrated in **Figure 1.b**, the RotaFormer head adopts the backbone architecture of the RotaFormer module for protein structure prediction with several modifications. The adapted head utilizes both 1D and 2D features as inputs and deploys 12 blocks of Sequence to Sequence (Seq2Seq) submodules and Pair to Pair (Pair2Pair) submodules. The Seq2Seq submodules feature Sequence Row Attention with Pair Bias, while the Pair2Pair submodules feature Triangle Multiplication and Triangle Attention. This is followed by a 5x5 kernel CNN to obtain the final probability map for base-pairing.

Benchmarking Domain-Specific DNNs. For the secondary structure domain-specific deep learning models, we obtained their released weights and tested their performance on TS0 under a unified pipeline. In the following we refer to the original train-validation-test split data from bpRNA-1m as TR0, VL0, and TS0. The SPOT-RNA model was trained on TR0 and then fine-tuned on a curated small high-resolution structure data from PDB, which is the only weight available. MXFold2 was trained on TR0 and validated on VL0 from bpRNA-1m. And for UFold, we utilized its released ufold_train.pt weight.

Post-Process. Typically, deep-learning methods output a 2D contact interaction map, which for certain models necessitates additional post-processing techniques to derive definitive base-pairs from the probability map. These methods often involve enforcing symmetry on the contact matrix. For instance, UFold screens out non-canonical base pairings and restricts sharp loops. In contrast, the RiNALMo-head leverages a greedy algorithm to iteratively establish base pairs with the highest pairing probabilities while excluding potential clashing pairs.

Metrics. It's worth noting that, as conformational dynamics exist in base pairing, which means that base pairs can fluctuate, the $(i\pm1, j)$, $(i, j\pm1)$ alternative states regarding the (i, j) base pair should also be considered correct in secondary structural prediction [12]. This is often reflected in the metric F1-shifted in secondary structure prediction task.

Protein-RNA Interaction

Dataset. From the Hela clip data provided by PrismNet, for Seq+SS model, the icSHAPE data are utilized, while for other models, only the RNA fragments sequences are taken as input. According to the original setting in PrismNet, 20% of the samples are taken as test set. Differently, we added train-validation split that randomly takes out 20% of the training set as validation set. The random seed is set the same across different models as default.

Feature Reduction. To optimize the utilization of language model features in PrismNet, we performed dimensionality reduction on the original high-dimensional features. Specifically, language features were reduced to 32 dimensions using a trainable linear layer. Similarly, one-hot sequence embeddings were transformed to 16 dimensions to maintain consistency in magnitude. Hyperparameters were aligned with PrismNet settings: the sequence-only model utilized a kernel height of 3, whereas models incorporating icSHAPE data or language model features utilized a kernel height of 5.

Mean Ribosome Loading

Dataset. The dataset used is a large-scale synthetic Human 5′ UTR library comprising 5′ UTR sequences with measured MRL values originally described by [35]. It includes human and random UTR sequences of varying lengths, with 100 sequences per length (from 25 to 100 nucleotides) selected based on deep read coverage. This sampling also created two evaluation datasets, Random7600 and Human7600, ensuring balanced representation across sequence lengths.

Head Architecture. The MRL prediction head is the same as the one used by RiNALMo, despite for RNA-FM feature input, the feature dimension of the first linear layer is 640 instead of 1280 for ProtRNA and RiNALMo.

Data availability

The code repository and pretrained model weights are available at https://github.com/roxie-zhang/ProtRNA

Acknowledgements

J.M. wants to thank the support from the Science and Technology Innovation Plan of Shanghai Science and Technology Commission (No. 23JS1400200), and the Research Fund for International Senior Scientists (No. W2431060). G.X. wants to thank the support from the National Natural Science Foundation of China (No. 32300535).

Author Contributions

R.Z. and G.X. jointly developed and trained the model, with R.Z. writing the primary scripts and G.X. refining the implementation. R.Z. also performed the analyses and drafted the manuscript. R.Z., B.M., and G.X. collaborated on downstream experiments. J.M. and G.X. supervised the research and designed the study. All authors contributed to project discussions and manuscript revisions.

References

1. Devlin, J., et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Naacl Hlt 2019), Vol. 1, 2019: p. 4171-4186.
2. Lin, Z.M., et al., *Evolutionary-scale prediction of atomic-level protein structure with a language model.* Science, 2023. **379**(6637): p. 1123-1130.
3. 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 of the United States of America, 2021. **118**(15).
4. Vig, J., et al., *Bertology meets biology: Interpreting attention in protein language models.* arXiv preprint arXiv:2006.15222, 2020.
5. Chen, J., et al., *Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions.* arXiv preprint arXiv:2204.00300, 2022.
6. *RNAcentral: a hub of information for non-coding RNA sequences.* Nucleic Acids Research, 2019. **47**(D1): p. D221-D229.
7. Penić, R.J., et al., *Rinalmo: General-purpose rna language models can generalize well on structure prediction tasks.* arXiv preprint arXiv:2403.00043, 2024.
8. Zhang, Y.K., et al., *Multiple sequence alignment-based RNA language model and its application to structural inference.* Nucleic Acids Research, 2024. **52**(1).
9. Johnsson, P., et al., *Evolutionary conservation of long non-coding RNAs; sequence, structure, function.* Biochimica et Biophysica Acta (BBA)-General Subjects, 2014. **1840**(3): p. 1063-1071.
10. Artetxe, M., S. Ruder, and D. Yogatama, *On the Cross-lingual Transferability of Monolingual Representations.* 58th Annual Meeting of the Association for Computational Linguistics (Acl 2020), 2020: p. 4623-4637.
11. Steinegger, M. and J. Söding, *MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.* Nature Biotechnology, 2017. **35**(11): p. 1026-1028.
12. Mathews, D.H., *How to benchmark RNA secondary structure prediction accuracy.* Methods, 2019. **162**: p. 60-67.
13. UniProt Consortium, T., *UniProt: the universal protein knowledgebase.* Nucleic acids research, 2018. **46**(5): p. 2699-2699.
14. Van der Maaten, L. and G. Hinton, *Visualizing data using t-SNE.* Journal of machine learning research, 2008. **9**(11).
15. Rogers, A., O. Kovaleva, and A. Rumshisky, *A primer in BERTology: What we know about how BERT works.* Transactions of the Association for Computational Linguistics, 2021. **8**: p. 842-866.
16. Vig, J. and Y. Belinkov, *Analyzing the structure of attention in a transformer language model.* arXiv preprint arXiv:1906.04284, 2019.
17. Clark, K., *What Does Bert Look At? An Analysis of Bert's Attention.* arXiv preprint arXiv:1906.04341, 2019.

18. Voita, E., et al., *Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned.* arXiv preprint arXiv:1905.09418, 2019.
19. Rao, R., et al., *Transformer protein language models are unsupervised structure learners.* bioRxiv, 2020: p. 2020.12.15.422761.
20. Kalvari, I., et al., *Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families.* Nucleic Acids Research, 2018. **46**(D1): p. D335-D342.
21. Nawrocki, E.P. and S.R. Eddy, *Infernal 1.1: 100-fold faster RNA homology searches.* Bioinformatics, 2013. **29**(22): p. 2933-2935.
22. Singh, J., et al., *RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning.* Nature Communications, 2019. **10**.
23. Fu, L.Y., et al., *UFold: fast and accurate RNA secondary structure prediction with deep learning.* Nucleic Acids Research, 2022. **50**(3).
24. Sato, K., M. Akiyama, and Y. Sakakibara, *RNA secondary structure prediction using deep learning with thermodynamic integration.* Nature Communications, 2021. **12**(1).
25. Xu, G., et al., *OPUS-Rota5: A highly accurate protein side-chain modeling method with 3D-Unet and RotaFormer.* Structure, 2024.
26. Reuter, J.S. and D.H. Mathews, *RNAstructure: software for RNA secondary structure prediction and analysis.* BMC bioinformatics, 2010. **11**: p. 1-9.
27. Al-Hashimi, H.M., *Turing, von Neumann, and the computational architecture of biological machines.* Proceedings of the National Academy of Sciences, 2023. **120**(25): p. e2220022120.
28. Gerstberger, S., M. Hafner, and T. Tuschl, *A census of human RNA-binding proteins.* Nature Reviews Genetics, 2014. **15**(12): p. 829-845.
29. Castello, A., et al., *Insights into RNA biology from an atlas of mammalian mRNA-binding proteins.* Cell, 2012. **149**(6): p. 1393-1406.
30. Sun, L., et al., *Predicting dynamic cellular protein–RNA interactions by deep learning using in vivo RNA structures.* Cell research, 2021. **31**(5): p. 495-516.
31. Kozak, M., *Structural features in eukaryotic mRNAs that modulate the initiation of translation.* Journal of Biological Chemistry, 1991. **266**(30): p. 19867-19870.
32. Araujo, P.R., et al., *Before it gets started: regulating translation at the 5′UTR.* International Journal of Genomics, 2012. **2012**(1): p. 475731.
33. Chu, Y., et al., *A 5′UTR language model for decoding untranslated regions of mRNA and function predictions.* Nature Machine Intelligence, 2024. **6**(4): p. 449-460.
34. Karollus, A., Ž. Avsec, and J. Gagneur, *Predicting mean ribosome load for 5'UTR of any length using deep learning.* PLoS computational biology, 2021. **17**(5): p. e1008982.
35. Sample, P.J., et al., *Human 5′UTR design and variant effect prediction from a massively parallel translation assay.* Nature biotechnology, 2019. **37**(7): p. 803-809.
36. Wang, Z., et al., *Cross-lingual alignment vs joint training: A comparative study and a simple unified framework.* arXiv preprint arXiv:1910.04708, 2019.
37. Conneau, A. and G. Lample, *Cross-lingual language model pretraining.* Advances in neural information processing systems, 2019. **32**.

38. Liu, Y.H., et al., *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* Information Systems Research, 2019.
39. Ren, Y., et al., *BEACON: Benchmark for Comprehensive RNA Tasks and Language Models.* arXiv preprint arXiv:2406.10391, 2024.