MiRformer: a dual-transformer-encoder framework for predicting microRNA-mRNA interactions from paired sequences

Jiayao Gu[1,2], Can (Sam) Chen[2], and Yue Li[1,2,3]

[1] School of Computer Science, McGill University, Montréal QC, Canada
[2] Mila - Quebec AI Institute, Montréal QC, Canada
[3] Correspondence to yueli@cs.mcgill.ca

Abstract. MicroRNAs (miRNAs) are small non-coding RNAs that regulate genes by binding to target messenger RNAs (mRNAs), causing them to degrade or suppressing their translation. Accurate prediction of miRNA–mRNA interactions is crucial for RNA therapeutics. Existing methods rely on handcrafted features, struggle to scale to kilobase-long mRNA sequences, or lack interpretability. We introduce **MiRformer**, a transformer framework designed to predict not only the binary miRNA–mRNA interaction but also the start and end location of the miRNA binding site in the mRNA sequence. MiRformer employs a dual-transformer encoder architecture to learn interaction patterns directly from raw miRNA-mRNA sequence pairs via the cross-attention between the miRNA-encoder and mRNA-encoder. To scale to long mRNA sequences, we leverage *sliding-window attention* mechanism. MiRformer achieves state-of-the-art performance across diverse miRNA–mRNA tasks, including binding prediction, target-site localization, and cleavage-site identification from Degradome sequencing data. The learned transformer attention are highly interpretable and reveals highly contrasting signals for the miRNA seed regions in 500-nt long mRNA sequences. We used MiRformer to simultaneously predict novel binding sites and cleavage sites in 13k miRNA-mRNA pairs and observed that the two types of sites tend to be close to each other, supporting miRNA-mediated degradation mechanism. Our code is available at https://github.com/li-lab-mcgill/miRformer.

Keywords: microRNA · Genome Language Model · Transformer · Post-transcriptional regulation.

1 Introduction

MicroRNAs (miRNAs) are small endogenous non-coding RNAs of approximately 22 nucleotides that regulate gene expression by binding to target messenger RNAs (mRNAs), leading to their degradation or translational repression [1]. Each miRNA can target hundreds of mRNAs, forming complex post-transcriptional regulatory networks that control diverse biological processes such as cell proliferation, differentiation, and apoptosis. Dysregulation of miRNA expression has been implicated in numerous diseases, including cancer and neurological disorders [2, 3]. Accurate prediction of miRNA–mRNA interactions is therefore crucial for understanding gene regulatory mechanisms and for advancing RNA-based therapeutics. However, identifying true target sites remains highly challenging, as many functional interactions involve non-canonical seed regions that deviate from perfect Watson–Crick-Franklin base pairing [4], complicating the interpretation of miRNA targeting rules and motivating the development of more accurate computational models.

Although experimental assays and next-generation sequencing approaches can map miRNA–mRNA interactions, they remain labor- and time-intensive, motivating the development of accurate in-silico predictors. Traditional heuristic tools such as TargetScan scan the 3′UTRs of target mRNAs for conserved seed matches across species [5]. While effective for canonical sites, these approaches rely on predefined alignment and conservation features, making them difficult to generalize to novel binding patterns or scale to kilobase-long sequences. To improve accuracy, deep learning frameworks such as DeepMirTar incorporates hundreds of handcrafted biological features—including seed match, free energy, and sequence stability—alongside sequence encodings [6]. However, these features are derived from limited experimental datasets and often fail to capture the full diversity of functional interactions. miRAW runs ViennaRNA on concatenated miRNA-mRNA sequences to obtain potential binding sites and relies on linear layers to identify true interacting sites [7]. ViennaRNA [8] predicts secondary structures with minimum free energy and base pairings with the highest probabilities, both are complementary features to raw RNA sequences which contribute to model instability. The miTAR framework is introduced as an advanced system utilizing a sophisticated hybrid deep

learning model [9]. This model integrates convolutional neural networks with recurrent neural networks to significantly enhance the capacity for recognizing patterns within biological data. However, this capability is specifically limited to analyzing short concatenated sequences of miRNA and mRNA, which presents a constraint in its application scope. REPRESS is one of the recent models that uses a convolution architecture to predict miRNA-mRNA interactions [10]. It feeds mRNA sequences into four blocks consisting of Residual blocks and convolution layers with increasing dilation rate. It then concatenates the outputs from all the blocks and feeds the concatenated output to two heads - one head predicting miRNA binding probability and the other predicting Degradome sequencing read counts at base-level resolution. It predicts miRNA targets output for human and mouse in 29 cell lines and tissues, and 10 cell lines and tissues for Degradome-seq outputs [10]. While predicting cell line and tissue-specific output is appealing, the model does not use miRNA sequence as input information and only predicts miRNA target sites abuntantly expressed in the targeted cell lines and tissue, which constraints the model from generalizing to the prediction of novel miRNA target sites. Overall, existing methods mainly suffer from two fundamental limitations: reliance on manually crafted features derived from limited experimental data, and poor scalability to kilobase-long mRNA sequences.

Recent advances in transformer models have revolutionized biological sequence modeling by capturing both global and local contextual relationships through self-attention mechanisms [11–18]. In particular, transformer-based frameworks in genomics have demonstrated strong generalization across diverse tasks by learning hierarchical regulatory patterns directly from raw nucleotide sequences. In this study, we present **MiRformer**, a hybrid convolution and transformer framework designed to simultaneously predict sequence-level miRNA-mRNA interaction and the nucleotide-level miRNA binding sites (i.e., seed regions). We make three technical contributions. First, we propose an effective dual-transformer-encoder architecture, where miRNA and mRNA sequences are processed independently through parallel transformer encoders. Second, to scale to long mRNA sequences, we leverage Longformer approach, which uses sliding window local self-attention to effectively model long-range dependencies with reduced computational cost compared to the full self-attention [19]. Third, to model miRNA-mRNA nucleotide-level interactions, we propose a novel sliding-window cross-attention and Log-Sum-Exponential (LSE) pooling layer, enabling clear focus on true seed regions even in long mRNA sequence context and therefore improving the model interpretabiblity. Through comprehensive benchmark using TargetScan and experimental Degradome-seq data, MiRformer demonstrates state-of-the-art performance across binding prediction, seed-region localization, and cleavage-site identification, consistently outperforming baseline models. Furthermore, MiRformer demonstrates strong zero-shot performance, identifying biologically plausible seed and cleavage sites without paired training examples.

2 Method

2.1 MiRformer model details

Convolution layers to extract initial sequence features All miRNA and mRNA sequences are tokenized at the nucleotide level. To enhance local continuity, the ribonucleotide tokens are first processed through two convolution layers with kernel sizes of 5 and 7, respectively (Fig. 1a). This design is inspired by RNAGenesis [20], which demonstrated that convolution-augmented embeddings improve token coherence.

Dual-transformer-encoder architecture with distinct self-attention mechanisms miRFormer uses two dedicated transformer encoders that are jointly trained end-to-end (Fig. 1a) with distinct attention mechanisms: **full self-attention** and **sliding-window attention** for miRNA and mRNA, respectively (Fig. 1b). Full self-attention encodes the complete miRNA sequence, allowing each nucleotide to attend to all other nucleotides. For mRNA sequences, we use sliding-window attention because it not only significantly reduces computational cost on modeling long sequences but also closely aligns with the biological mechanism of miRNA regulation where part of the 22-nt miRNA binds to small segment of a much longer mRNA sequence. Inspired by Longformer [21], sliding-window attention encodes the mRNA sequence using a window of size $2w$ with $w = 20$nt that slides across the mRNA sequence. The window of $2w$ width is centered on every nucleotide of mRNA sequence, allowing it to only attend to nucleotides inside the window. We segment the query and key mRNA embedding into overlapping chunks of $2w$, compute attention within chunks, and merge chunks into the attention matrix without duplicate chunks. All matrix calculation is fully vectorized to reduce the time complexity from $O(L^2)$ with standard transformer attention to $O(wL)$ for L mRNA length.

Cross-attention Once we have the miRNA embedding and mRNA embedding, we need to design a sliding window cross-attention mechanism to let miRNA sequence efficiently attend mRNA sequence, thereby

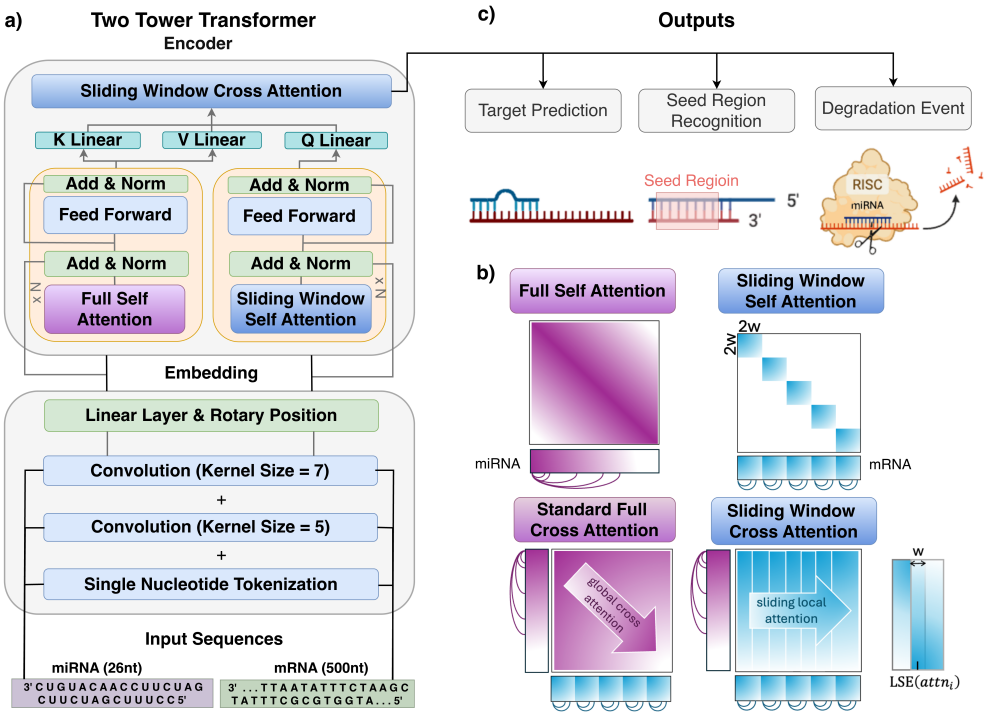

Fig. 1: MiRformer overview a) MiRformer leverages convolutional tokenization layers to extract seqeunce features from miRNA and mRNA, and encodes the features via a Dual-Transformer-Encoder Architecture (DTEA). The resulting embeddings are then fed to two dedicated transformer encoders for process miRNA and mRNA. The two embeddings are then fused via cross-attention that emulates the mRNA-mRNA nucleotide-level sequence recognition. b) Attention mechanisms. Full self attention encodes miRNA sequences by attending every base of the input miRNA sequence to itself and other bases. Sliding-window attention encodes mRNA sequences by attending only bases within a $2w$ window. Cross attention fuses both miRNA and mRNA sequence embeddings by scanning a $2w$ window across the mRNA embedding. c) Training tasks: target prediction, seed region recognition and degradation event prediction.

mimicking the biological miRNA-mRNA ribonucleotide-level hybridization. We note that self-attention is a square matrix whereas cross-attention between miRNA and mRNA is a rectangular matrix, which requires a different design to merge the overlapping sliding windows (Fig. 1b). Let $C = \frac{L}{w} - 1$ be the number of sliding-window chunks for an mRNA of length L with window width $w = 20$ nt. Each chunk is indexed by $c \in \{1, \ldots, C\}$ and each position inside a chunk is indexed by $i = c \times w + u$, where u is the within-chunk offset. The overlapping length between adjacent chunks is $o = \frac{L}{w} - 2w$. Before chunking, the cross-attention matrix is $z \in \mathbb{R}^{L_q, L_k}$; after merging overlapping windows, it becomes $z \in \mathbb{R}^{L_q+o, L_k}$. To merge signals from all chunks $C(i)$ covering the same position i, we apply LSE pooling: $z_{i,l_k} = \log \sum_{c \in C(i)} \exp(z^c_{i,l_k})$. We use LSE pooling instead of mean pooling because LSE preserves strong local alignment signals by approximating a max operator, while still aggregating information across windows. This prevents true binding-site signals from being diluted by low-attention regions, which frequently occurs when merging overlapping windows in long mRNA sequences. Finally, attention is normalized either along miRNA positions $p_{i,\cdot} = \frac{\exp(z_{i,j})}{\sum_j \exp(z_{i,j})}$ or along mRNA positions $p_{\cdot,j} = \frac{\exp(z_{i,j})}{\sum_i \exp(z_{i,j})}$, as our experiments explored (Fig. 5).

2.2 Datasets

TargetScan dataset Positive samples were constructed from TargetScan (version 7.2) predictions with Context++ scores above a specified threshold (top 20% by percentile). For each positive miRNA-mRNA pair, we extracted a window around the seed region. Negative samples were generated by: 1. Shuffling miRNA-mRNA pairs to create non-binding pairs, and 2. Extracting segments from mRNA sequences that do not contain complementary seed matches with the miRNA. The window size in this context were set to

be either 30-nt or 500-nt. In total, we curated human and mouse miRNA–mRNA pairs and constructed a dataset of 560k examples, comprising 280k positive pairs containing positive miRNA target sites and 280k negative pairs lacking predicted target sites.

Seed position randomization To prevent model from taking short-cut (i.e., memorizing the location of the seed region), seed regions were randomly positioned within the 30-nt and 500-nt segment. The left extension was randomly chosen between total extension and seed start, where total extension = window length - seed length. The right extension is total extension - left extension. If the right extension exceeds the mRNA sequence length, the window is left-adjusted accordingly. Similarly for the left extension.

Degradome-seq dataset Paired mRNA-miRNA and experimental cleavage sites were collected from Starbase v2.0 [22]. The 3'UTR sequences and gene coordinates were extracted from GENCODE v19 annotations [23] for GRCh37 (hg19). 500-nt windows around the cleavage sites were extracted and used to construct the training and testing data. In total, we collected 210k miRNA-mRNA pairs with observed cleavage sites.

2.3 Prediction Tasks

The 3 tasks are illustrated in Fig. 1c and described below.

Target prediction MiRformer was trained to classify each miRNA–mRNA pair as interacting (positive) or non-interacting (negative). Specifically, the output embeddings from the last cross-attention layer are fed to into a linear classifier, which outputs a sequence-level sigmoid probability.

Seed-region recognition MiRformer was trained to predict the start and end positions of the binding region within mRNA sequences. The nucleotide-level embedding is fed to a linear layer to give softmax probabilities at every position of either 1) being the start of a seed region or 2) being the end of a seed region.

For both tasks, MiRformer was trained on TargetScan Human and Mouse v7.2 [5].

Degradation event prediction We obtained Degradome-seq experimental data from Starbase v2.0 [22] and seqeunces from GENCODE v19 [23] cleavage site prediction.

2.4 Objective functions

MiRformer minimizes a combined loss function for task 1 and 2 described above:

$$\mathcal{L} = \alpha_1 \mathcal{L}_1 + \alpha_2 \mathcal{L}_2$$

where $\mathcal{L}_1$ is the binary cross-entropy loss for target prediction:

$$\mathcal{L}_1 = -y \log(x) - (1-y)\log(1-x),\ x = \frac{1}{1+\exp(-z)}$$

and z is the sequence-level embedding; $\mathcal{L}_2$ is the cross-entropy loss for seed-region recognition:

$$\mathcal{L}_2 = -\sum_{c}^{C} y_c \log \frac{\exp(z_c)}{\sum_i^C \exp(z_i)}$$

where z_c is the token embedding of nucleotide c. We set $\alpha_1 = 1$ and $\alpha_2 = 0.75$. Note that $\mathcal{L}_2$ is computed only for positive samples.

After training on TargetScan data, MiRformer was continuously trained on Degradome-seq data to predict cleavage sites within mRNA sequences. We transformed cleavage site coordinates into Gaussian-smoothed labels of size $y = \exp\left(\frac{(i-\mu)^2}{2\sigma^2}\right) \in \mathbb{R}^L$, where i is the query position and μ is the cleavage site position, L is mRNA length, and standard deviation $\sigma = 3.0$. We define the loss as a Gaussian-weighted binary cross entropy (BCE) loss. Let $\mathcal{L}_i = -y_i \log(\hat{y_i}) - (1-y_i)\log(1-\hat{y_i})$ and $w_i = 1 + (\lambda - 1)y_i$ where $\lambda = [1, 2]$ is a weighting factor to enhance the gradient at $y_i = 1$. The weighted BCE loss is $\sum_{i=1}^{L} \frac{w_i \mathcal{L}_i}{\sum_{i=1}^{L} w_i}$.

2.5 In-silico mutagenesis

We perturbed every base and recorded the change in probability and the change in attention score. Specifically, each base was perturbed 3 times from the original base to the 3 alternative bases. We then averaged the change over the 3 perturbations. Formally, x_i^u denotes the substituted base u at position $i = 1 \cdots L$,:

$$s_1(x_i^u) = M(x_i) - M(x_i^u); \quad s_2(x_i^u) = \text{Attention}(x_i) - \text{Attention}(x_i^u)$$

where $M(x_i)$ denotes the model's prediction probability for the original sequence, and $M(x_i^u)$ denotes the prediction probability after replacing base x_i with base u. And the final perturbation score is $s_f = \frac{1}{3}\sum_u s(x_i^u), u \in \{A, C, T, G\}$ for $f \in \{1, 2\}$.

2.6 Baseline methods

For TargetScan, we downloaded the software from TargetScan website `https://www.targetscan.org`. We followed the instructions to run TargetScan on custom data. Pairs of miRNA-mRNA that have predicted seed regions were considered positive, otherwise negative. For miTAR, since it was only trained on 66 nt miRNA-mRNA concatenated length using 24-nt miRNA, we scanned the 500 nt mRNA by a 40-nt window and predicted seed regions within the window. For target prediction, we ran miTAR and scanned every 40-nt window and selected the window with the highest probability. For seed region prediction, we added a span head that was a copy of the BiLSTM inside the classification head and also scanned every 40-nt window for seed regions. The window having the highest sum of top-k ($k = 8$) probability was evaluated on seed start and end prediction. For REPRESS, the software requires the input sequence length to be 12.5k and pads the input to the same maximum length. Its outputs include probability of binding for 29 cell lines at each base pair as well as read counts of cleavage sites for 10 cells lines. We evaluated the outputs as described in [10] - average pooled both the prediction and labels within the predefined window lengths for each cell line and divided the output into windows.

3 Results

3.1 Hybrid embedding improves model performance as well as interpretability

As a proof-of-concept, we first evaluated the benefits of having the hybrid tokenization on short 30-nt and compared its performance with the one without convolution layer (Fig. 2). Model trained with convolutional tokenization demonstrated better performance than the ablated counterparts (Fig. 2a). We investigated model's interpretability by showing cross attention scores with and without convolution layers. We observed that ablated model attended to the boundaries of seed region instead of the nucleotides inside the seed (Fig

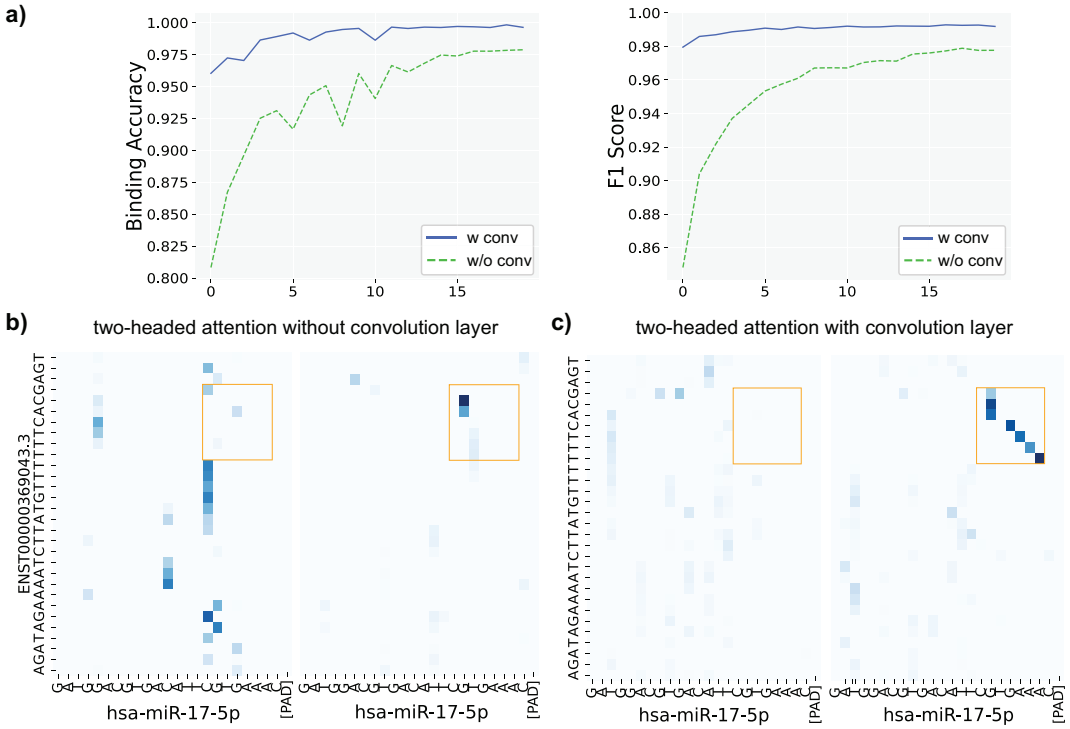

Fig. 2: Proof-of-concept analysis on 30-nt short mRNA sequences. a) Model performance comparison between with and without convolutional feature extraction. b) Heatmap produced by model trained without convolution feature embeddings. c) Heatmap produced by model with convolution feature embeddings. For both b and c, cross attention weights from two attention heads are shown as heatmaps produced by a dual-transformer-encoder architecture trained on both binding labels and seed span positions. Complimentary seed region within mRNA (y-axis) and miRNA (x-axis) are highlighted in box.

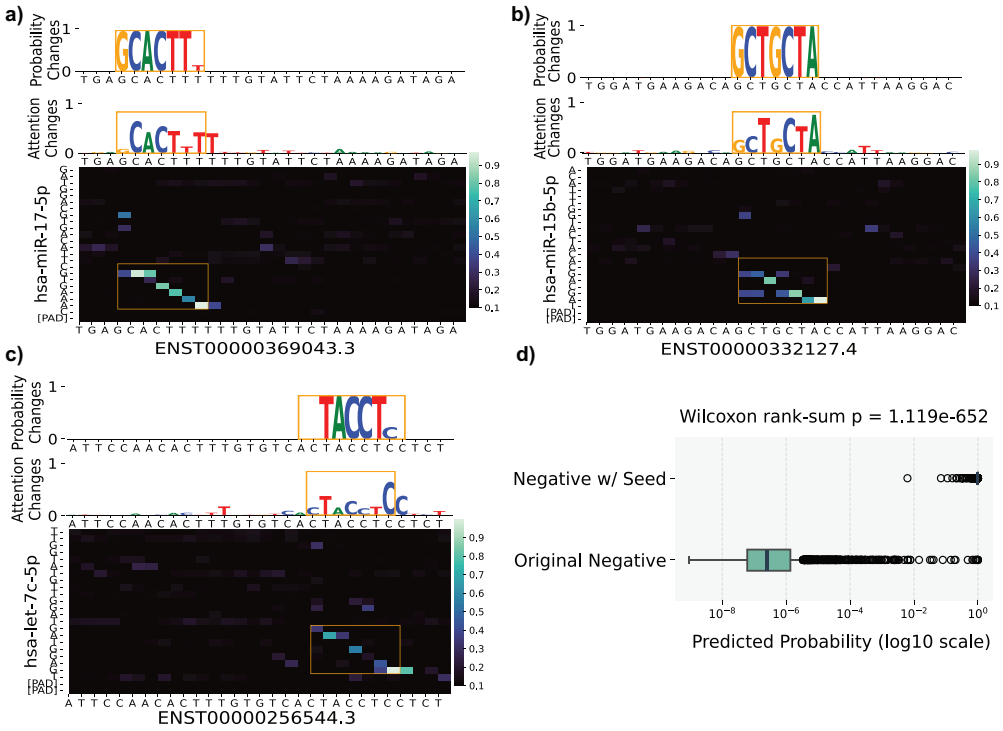

Fig. 3: In-Silico Mutagenesis (ISM) analysis of 30-nt mRNA sequences. (a-c) ISM and attention visualization. We perturbed every base and recorded the change in target prediction probability and the changes in cross attention matrix at each base. For each of the 3 examples, the top two tracks display the probability and attention change at each base. The bottom heatmap displays the cross-attention between miRNA and mRNA. d) Boxplot showing predicted probabilities on Original Negative – a set of negative miRNA-mRNA pairs without seeds, and Negative w/ Seed - Original Negative with inserted seed regions.

2b). In contrast, MiRformer with the convolution layer produced more prominent cross attention score at not only the boundaries of the seed region but also the nucleotides inside the seed region (Fig. 2c).

To further investigate model's interpretability on single-base level, we performed in-silico mutagenesis (ISM) and averaged probabilities at every base of all perturbations (Section 2.5). We showed change in target prediction scores and change in cross attention scores after ISM for randomly selected examples. The changes were the highest within seed regions (Fig. 3a-c). This provides evidence that our model predicts miRNA-mRNA bindings by identifying the seed regions. We also tested model's ability on predicting targets on negative miRNA-mRNA pairs with artifically inserted seed regions (Fig. 3e). We observed a significant difference in target prediction scores between the original negative miRNA-mRNA pairs and the negative pairs with inserted seeds. This further shows that our model is able to predict miRNA targets based on true seed regions. Therefore, the convolutional layers extract useful local features for more global transformer attention layers and improve model performance on both target prediction and attention interpretability.

3.2 Modeling long mRNA sequences

We experimented the benefits of the proposed sliding-window cross attention and LSE (Log-Sum-Exponential) pooling technique in modeling the interaction between miRNA and longer mRNA sequence segments of 500-nt in length. We compared 7 model variants with different combinations of mean and LSE pooling on target prediction and seed region recognition (Fig. 4). First, we observed an improved model performance with LSE pooling compared to mean pooling. Specifically, LSE pooling improved seed recognition by 0.20 on F1 score compared to mean pooling. Notably, model using just the leading CLS token as input for target prediction reduced accuracy by 0.40 (Fig. 4), implying that CLS alone cannot capture sufficient sequence information. Models with LSE at the sliding-window cross attention layer performed on par with model with global cross attention on target prediction. Both achieving 99% classification accuracy. The former conferred superior

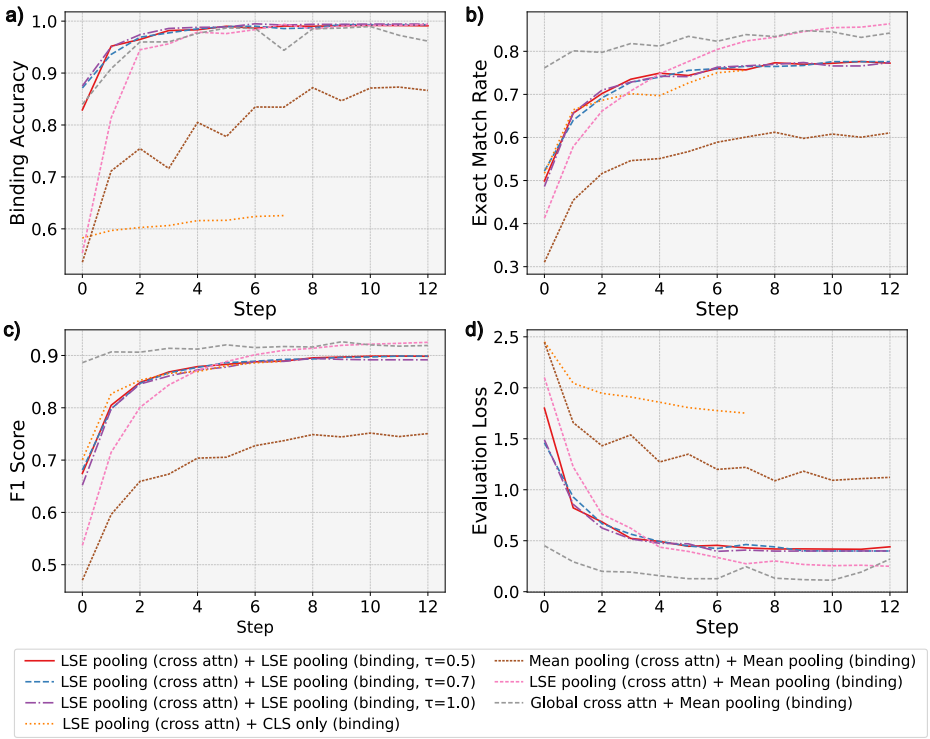

Fig. 4: Pooling methods comparison. We experimented different temperature values τ when applying LSE pooling on window embedding as well as only using the CLS token prepended to the mRNA sequence for target prediction. We monitored the validation performance as a function of training epochs. a) Prediction accuracy on miRNA-mRNA interaction. b) Prediction accuracy of predicted seed start and seed end. c) F1 score measuring overlapping regions between predicted seed regions and true seed regions. d) Loss of each model when evaluated on a separate validation set for target and seed predictions.

accuracy on seed region recognition than the latter - 0.86 vs 0.85 (Fig. 4c) while maintaining similar seed F1 score 0.91 (Fig. 4d). In conclusion, we showed that LSE improved all metrics compared to mean pooling and performed at least on-par compared with global cross attention. We reasoned that the improved performance by LSE compared to mean pooling could be attributed to LSE successfully de-convoluting positive signals from negative signals when merging windows during sliding-window cross attention.

3.3 Attention interpretability

To verify that our model truly learned seed regions in 500-nt mRNA sequence context, we examined the final cross-attention scores from the sliding-window cross attention layer in similar way as for the 30-nt analysis (Section 3.1). We also conducted ISM on mRNA sequence to investigated the importance of each base (Section 2.5). In a selected example, we observed that miRformer focused on seed regions both in cross-attention scores and signals from ISM tests (Fig. 5). We tried two ways to normalize cross attention: one was normalizing by mRNA length (Fig. 5a) and the other was normalizing by miRNA length (Fig. 5b). Normalizing by mRNA (miRNA) length was analogous to inquiring which ribonucleotides inside the mRNA (miRNA) sequence does the miRNA (mRNA) attends to the most. We observed much clearer pattern at the boundaries of the seed region when normalizing by miRNA (Fig. 5b). Nevertherless, attention heatmap pattern is much clearer in seed region compared to non-seed region when normalizing by mRNA (Fig. 5a), which underscored the importance of our sliding-window attention and LSE pooling on longer sequences. In contrast, mean pooling produces much noisier signals (Fig. 5c). Although global attention produced high prediction performance (Fig. 4), the resulting attention is not interpretable (Fig. 5d). In contrast, we observed a significant difference in average cross-attention scores between seed and non-seed regions for all sliding window approaches (Fig. 5a-c). However, the method of pooling proved critical. While the cross attention + mean probability approach (Fig. 5c) did yield a significant difference, the statistical separation

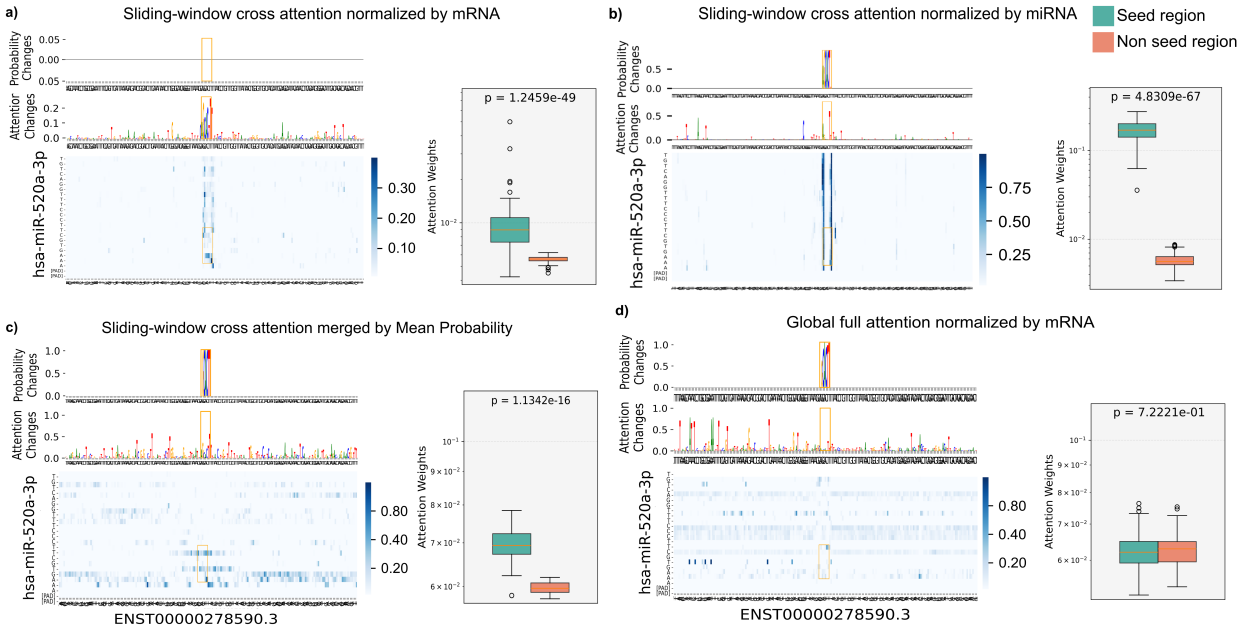

Fig. 5: Model interpretability through cross attention heatmaps on a 500-nt mRNA sequence. For each panel, top track displays the change in probability at the single nucleotide resolution after ISM; middle track shows the change in attention after ISM; bottom track displays cross attention between miRNA and mRNA at the nucleotide resolution without ISM perturbation. Boxplots show the average attention weights distributions of seed and non-seed regions. P-values were calculated by Wilcoxon rank-sum test on 200 miRNA-mRNA pairs between base pairs in the seed region and base pairs outside the seed region. a) Sliding-window cross attention + LSE pooling and normalized by mRNA. b) Sliding-window cross attention + LSE pooling and normalized by miRNA. c) Sliding-window cross attention + mean pooling and normalized by mRNA. d) Global cross attention and normalized by mRNA.

was orders of magnitude lower ($p \approx 10^{-16}$) compared to the cross attention + LSE approaches ($p \approx 10^{-49}$ to $p \approx 10^{-67}$). This suggests that mean pooling dilutes the strong signals from seed regions by averaging them with the surrounding noise. Therefore, MiRformer utilizing sliding window cross-attention with LSE pooling provides the optimal balance - conferring highly interpretable attention without compromising performance.

3.4 MiRformer achieves state-of-the-art performance on all prediction tasks

We evaluated MiRformer against three state-of-the-art models (TargetScan, REPRESS, miTAR) for 3 tasks - target prediction, seed region recognition and experimental degradation event prediction (Section 2.3) and measured models' performance by 6 metrics (Fig. 6). For the first two tasks, we used the TargetScan Human and Mouse (version 7.2) as the gold standard labels, which were directly downloaded from the TargetScan database and originally derived by using the entire mRNA 3'UTR sequences. The local TargetScan algorithm was tested on only 500-nt mRNA 3'UTR segments and so as all other computational methods. Accurately predicting TargetScan gold standard labels are useful because of (1) the much more efficient forward-pass inference algorithm compared to the traditional multiple sequence alignments to compute conservation scores followed by rule-based scoring over the entire 3'UTR sequences and (2) the fact that the deep learning method can then be used to scan all miRNA-mRNA sequence pairs beyond canonical seed match rules as we demonstrated in Section 3.5. MiRformer outperformed three other models on all 6 metrics (Fig. 6 top row). Notably, MiRformer exceeded the second best model on predicting Degradome-seq cleavage sites for Hit at 5-nt accuracy by 0.348 and on Hit at 3-nt by 0.339.

3.5 Joint inference of miRNA-binding sites and cleavage sites

While Degradome-seq provides experimental evidence on miRNA-mRNA interactions via observed cleavage sites within 3'UTR, it does not produce the exact miRNA binding sites. TargetScan generates computational seed-match predictions, but only a fraction of those were supported by Degradome-seq data. We sought an

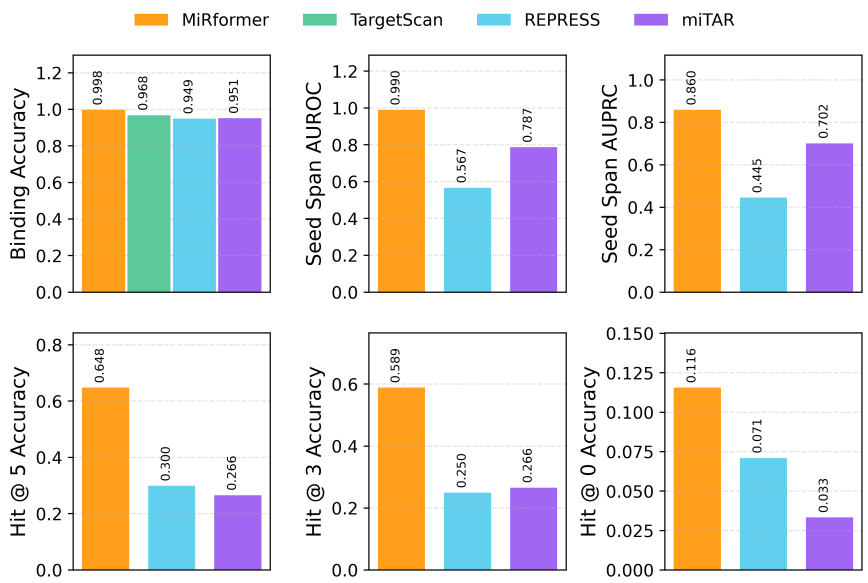

Fig. 6: Benchmark on seed region and cleavage site prediction. Binding Accuracy evaluates classification of miRNA-mRNA interactions. Seed Span AUROC and AUPRC evaluate nucleotide-level prediction for inside the seed region. Hit @ k Accuracy measure accuracies of predicting the cleavage site within k-nt distance to the true cleavage site in the target mRNA sequence.

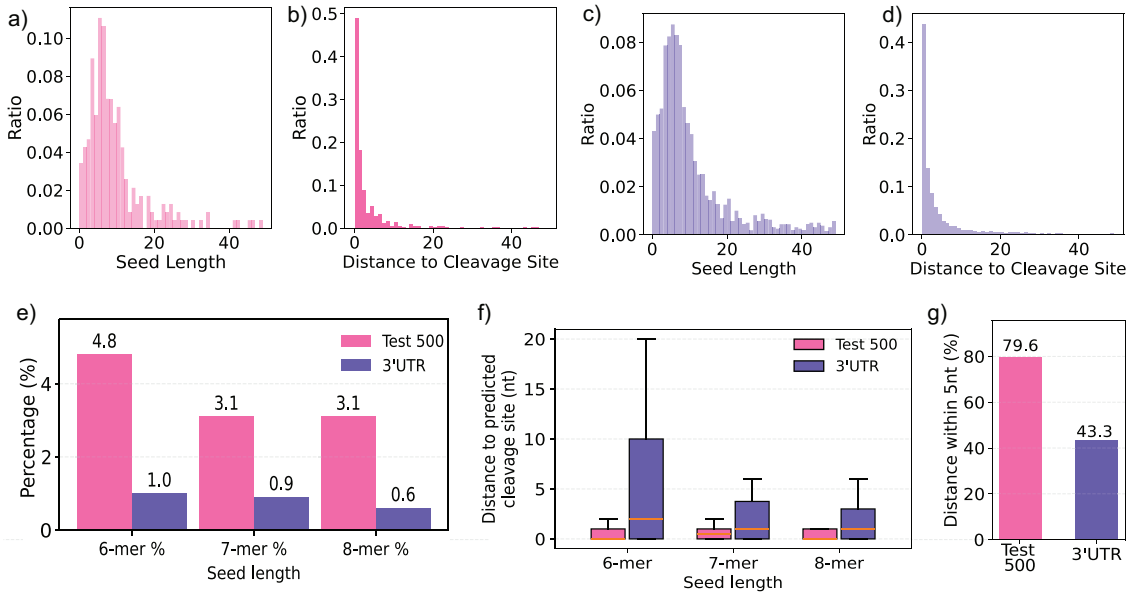

Fig. 7: Predicted seed regions and cleavage sites. a) Length distribution of the predicted seed regions on the Test 500 dataset (i.e., 500-nt mRNA 3'UTR segments). b) Distribution of the distance between the predicted seed regions and the predicted cleavage sites on Test 500 dataset. c,d) Same as a and d but for the predictions on the full-length 3'UTRs. e) Percentage of the predicted cleavage sites that fall within 5-nt to the predicted seed regions. f) k-mer distributions ($k \in \{6, 7, 8\}$) of the predicted seed region for Test 500-nt and full-length 3'UTR sequences. g) Percentage of the predicted cleavage sites that fall within 5-nt to seed regions for 500-nt and 3'UTRs.

approach to to simultaneously predict both seed region span and cleavage sites. Specifically, we first trained MiRformer on TargetScan data and then continuously trained it on Degradome-seq data for predicting cleavage sites (Section 2.4). We tested MiRformer on two datasets: 1. Test 500 - a TargetScan subset containing 500-nt mRNA paired with miRNA unseen during training. 2. UTR - complete full-length 3'UTR of mRNAs of up to 25,000 nt (Fig. 7). For the 500-nt mRNA segments, most of the predicted seed regions span 0 to 10 nt (Fig. 7a) and fell within 0 to 10 nt distance from the predicted cleavage sites (Fig. 7b). About 4.8% of them were 6-mer, 3.1% were 7-mer and 3.1% were 8-mer (Fig. 7e). Also, most of the predicted 6, 7 or 8-mer seed regions fell within only 2.5 nt from the predicted cleave sites (Fig. 7f). For the full-length 3'UTR sequences, 3'UTR sequences were padded to the longest in a one batch to predict both the seed regions and the cleavage sites. Most of the predicted seed regions were within 0 to 20 nt (Fig. 7c) and located within 10 nt from the predicted cleavage sites (Fig. 7d). Over 1.0% of all predicted seed regions were 6-mers, 0.9% were 7-mer and 0.6% were 8-mer (Fig. 7e). Lastly, 79.6% and 43.4% of predicted cleavage sites were within 5 nt distance to seed regions within 500-nt and full-length 3'UTR, respectively (Fig. 7g). Together, we demonstrated that MiRformer was capable of simultaneously predicting reasonable seed regions and cleavage sites despite not being previously trained on mRNAs containing paired seed regions and cleavage sites.

4 Discussion

Characterizing sequence features that dictate the miRNA-mRNA interactome is a long-standing challenge . Traditional methods such as TargetScan [24] incorporate handcrafted features such as canonical 6-8 mer seed regions and evolutionary conservation but suffer from many exceptions. Existing deep learning methods using convolutional neural networks struggle to capture the nucleotide-resolution miRNA-mRNA interaction in long sequence contexts. Transformer-based language models have proven successes in modeling biological sequences by capturing both global and local contextual information through attention learning mechanism. In this paper, we developed MiRformer to jointly learn miRNA and mRNA sequence features through two dedicated encoders with distinct self-attention mechanisms and then learn their interactions through sliding-window cross attention and LSE pooling, which scale to long mRNA sequences to both reduce computing cost and improve model interpretability. Our benchmarks demonstrated that miRFormer achieved accurate and interpretable predictions across binding, seed span, and cleavage tasks.

One caveat of our approach is that MiRformer has only mainly been trained on canonical seed regions derived from TargetScan dataset, which may limit its generalizability to recognizing non-canonical seed regions. We are planning to train MiRformer on a compendium of experimentally validated data [22, 25]. Recently, genome language models conditioned on whole genome alignment information achieve state-of-the-art performance on variant prediction on human genome [26]. Therefore, one potential direction is to integrate cross-species multiple sequence alignments to find conserved target sites and improves detection of non-canonical seed regions.

Looking forward, future models could also incorporate multiomic contexts such as RNA secondary structures [9], single-cell gene expression data [27], and miRNA synergistic network [28, 29] to further generalize beyond canonical binding rules. Moreover, foundation-scale RNA language models, pretrained across diverse transcriptomes and perturbation datasets, may enable zero-shot prediction of regulatory effects for novel miRNAs or disease-associated variants [30]. These directions promise to transform miRNA research from feature engineering to fully data-driven reasoning, guiding the discovery of new miRNA functions and helping RNA therapy designs. We believe that MiRformer is a milestone for us to acheive these exciting goals in the near future.

5 Acknowledgements

J.G. was supported by DNA2RNA Scholarship in promoting innovative RNA therapies. Y.L. is supported by Canada Research Chair (Tier 2) in Machine Learning for Genomics and Healthcare (CRC-2021-00547) and CIHR Project Grant (202503PJT-540722-GMX-CFAA-255237). Special thanks to Keren Zhou who kindly provided access to Degradome sequencing data.

References

1. David P Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297, 2004.
2. Bo Wang, Shu-hao Hsu, Xinmei Wang, Huban Kutay, Hemant Kumar Bid, Jianhua Yu, Ramesh K Ganju, Samson T Jacob, Mariia Yuneva, and Kalpana Ghoshal. Reciprocal regulation of microrna-122 and c-myc in hepatocellular cancer: role of e2f1 and transcription factor dimerization partner 2. *Hepatology*, 59(2):555–566, 2014.
3. Han Han, Dan Sun, Wenjuan Li, Hongxing Shen, Yahui Zhu, Chen Li, Yuxing Chen, Longfeng Lu, Wenhua Li, Jinxiang Zhang, et al. A c-myc-microrna functional feedback loop affects hepatocarcinogenesis. *Hepatology*, 57(6):2378–2389, 2013.
4. Vaclav Hejret, Nandan Mysore Varadarajan, Eva Klimentova, Katarina Gresova, Ilektra-Chara Giassa, Stepanka Vanacova, and Panagiotis Alexiou. Analysis of chimeric reads characterises the diverse targetome of AGO2-mediated regulation. *Scientific Reports*, 13(1):22895, 2023.
5. Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. Predicting effective microrna target sites in mammalian mrnas. *elife*, 4:e05005, 2015.
6. Ming Wen, Peisheng Cong, Zhimin Zhang, Hongmei Lu, and Tonghua Li. Deepmirtar: a deep-learning approach for predicting human mirna targets. *Bioinformatics*, 34(22):3781–3787, 2018.
7. Albert Pla, Xiangfu Zhong, and Simon Rayner. miRAW: A deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts. 14(7):e1006185.
8. Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA package 2.0. 6(1):26.
9. Tongjun Gu, Xiwu Zhao, William Bradley Barbazuk, and Ji-Hyun Lee. mitar: a hybrid deep learning-based approach for predicting mirna targets. *BMC bioinformatics*, 22:1–16, 2021.
10. Bhargav Kanuparthi, Sara E Pour, Scott D Findlay, Omar Wagih, Jahir M Gutierrez, Rory Gao, Jeff Wintersinger, Junru Lin, Martino Gabra, Emma Bohn, et al. Sequence based prediction of cell type specific microrna binding and mrna degradation for therapeutic discovery. *bioRxiv*, pages 2025–05, 2025.
11. Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.
12. Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. Nucleotide transformer: building and evaluating robust foundation models for human genomics. 22(2):287–297.
13. Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. 18(10):1196–1203.
14. Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brixi, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. Sequence modeling and design from molecular to genome scale with evo. 386(6723):eado9336, 2024.
15. Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36, 2024.
16. Sawan Patel, Fred Zhangzhi Peng, Keith Fraser, Adam D. Friedman, Pranam Chatterjee, and Sherwood Yao. EvoFlow-RNA: Generating and representing non-coding RNA with a language model. 2025.
17. Ning Wang, Jiang Bian, Yuchen Li, Xuhong Li, Shahid Mumtaz, Linghe Kong, and Haoyi Xiong. Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning. 6(5):548–557.
18. Manato Akiyama and Yasubumi Sakakibara. Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. 4(1):lqac012, 2022.
19. Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. (arXiv:2004.05150), 2020.
20. Zaixi Zhang, Ruofan Jin, Linlin Chao, Guangxue Xu, Yikun Zhang, Guowei Zhou, Di Yin, Yingqing Guo, Yaqi Fu, Yukang Yang, Kaixuan Huang, Xiaotong Wang, Junze Zhang, Yujie Yang, Qirong Yang, Ziyao Xu, Weinan E, Ruhong Zhou, Xiaoming Zhang, Mengdi Wang, and Le Cong. RNAGenesis: A generalist foundation model for functional RNA therapeutics.
21. Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.
22. Jun-Hao Li, Shun Liu, Hui Zhou, Liang-Hu Qu, and Jian-Hua Yang. starbase v2. 0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data. *Nucleic acids research*, 42(D1):D92–D97, 2014.

23. Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774, 2012.
24. Benjamin P Lewis, I-hung Shih, Matthew W Jones-Rhoades, David P Bartel, and Christopher B Burge. Prediction of mammalian microrna targets. *Cell*, 115(7):787–798, 2003.
25. Hsi-Yuan Huang, Yang-Chi-Dung Lin, Shidong Cui, Yixian Huang, Yun Tang, Jiatong Xu, Jiayang Bao, Yulin Li, Jia Wen, Huali Zuo, et al. mirtarbase update 2022: an informative resource for experimentally validated mirna–target interactions. *Nucleic acids research*, 50(D1):D222–D230, 2022.
26. Gonzalo Benegas, Carlos Albors, Alan J. Aw, Chengzhong Ye, and Yun S. Song. A DNA language model based on multispecies alignment predicts the effects of genome-wide variants. 2025.
27. Sarah M Hücker, Tobias Fehlmann, Christian Werno, Kathrin Weidele, Florian Lüke, Anke Schlenska-Lange, Christoph A Klein, Andreas Keller, and Stefan Kirsch. Single-cell microrna sequencing method comparison and application to cell lines and circulating lung tumor cells. *Nature communications*, 12(1):4316, 2021.
28. Yue Li, Cheng Liang, Ka-Chun Wong, Jiawei Luo, and Zhaolei Zhang. Mirsynergy: detecting synergistic mirna regulatory modules by overlapping neighbourhood expansion. *Bioinformatics*, 30(18):2627–2635, 2014.
29. Pei Liu, Ying Liu, Jiawei Luo, and Yue Li. Mirgraph: A hybrid deep learning approach to identify microrna-target interactions by integrating heterogeneous regulatory network and genomic sequences. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1028–1035. IEEE, 2024.
30. Pei Liu, Xiao Liang, Yue Li, and Jiawei Luo. Convntc: Convolutional neural tensor completion for predicting the disease-related mirna pairs and cell-related drug pairs. *bioRxiv*, pages 2024–10, 2024.