MiRformer: A Unified Generative Framework for mRNA-Conditioned miRNA Synthesis and Interaction Prediction

Jiayao Gu
School of Computer Science
McGill University
Mila - Quebec AI Institute
jia.gu@mail.mcgill.ca

Can Chen
Mila - Quebec AI Institute
can.chen@mila.quebec

Yue Li
School of Computer Science
McGill University
Mila - Quebec AI Institute
Correspondence: yue.li@cs.mcgill.ca

Abstract

MicroRNAs (miRNAs) are small non-coding RNAs that regulate gene expression by binding to target messenger RNAs (mRNAs), leading to mRNA degradation or translational repression. Accurate prediction of miRNA–mRNA interactions is critical for understanding post-transcriptional regulation and enabling RNA therapeutics. However, existing computational methods rely on handcrafted or indirect features, struggle to scale to kilobase-long mRNA sequences, or provide limited interpretability. We present MiRformer, a unified generative framework that models miRNA–mRNA interactions, pinpoints miRNA binding sites within long mRNA sequences, and synthesizes mRNA-conditioned miRNA sequences. MiRformer employs a dual-transformer encoder architecture to learn interaction patterns directly from raw miRNA–mRNA sequence pairs. To efficiently model long mRNA contexts, we incorporate a sliding-window attention mechanism that scales to kilobase-length sequences while preserving nucleotide-level resolution. MiRformer employees the pre-trained MiRformer encoder and trains a decoder to auregressively generate miRNA. MiRformer achieves state-of-the-art performance across multiple miRNA–mRNA prediction tasks, including interaction prediction, binding-site localization, and cleavage-site identification from experimental Human Degradome-seq data. Across all 500-nt mRNAs, MiRformer produces 5712 target-specific miRNA and 99.30% of these contain canonical seed regions. Beyond accuracy, MiRformer provides strong interpretability: attention patterns consistently highlight miRNA seed regions within 500-nt mRNA windows, revealing clear and biologically meaningful interaction signals. Our results demonstrate that MiRformer offers a scalable, accurate, and interpretable approach for modeling miRNA–mRNA interactions and synthesizing biologically plausible target-specific miRNA sequences. Code is available at https://github.com/li-lab-mcgill/miRformer.

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 (Bartel, 2004). 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 (Wang et al., 2014). Dysregulation of miRNA expression has been implicated in numerous diseases, including cancer and neurological disorders (Wang et al., 2014; Han et al., 2013). Accurate prediction of miRNA–mRNA interactions is crucial for understanding gene regulatory mechanisms and for advancing RNA-based therapeutics, 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 (Agarwal et al., 2015). 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, DeepMirTar and miRAW incorporate handcrafted biological features—including minimum free energy, base pairing probability and sequence stability —alongside raw sequences (Wen et al., 2018; Pla et al., 2018), but its reliance on limited experimental data and sensitivity to small perturbations to input sequence restrict model's generalizability. The miTAR framework integrates convolutional neural networks with recurrent neural networks to significantly enhance the capacity for recognizing patterns within biological data (Gu et al., 2021). However, this capability is specifically limited to analyzing short concatenated sequences of miRNA and mRNA, which presents a constraint in its application scope. REPRESS consists of a stack of residual connections between convolutional layers that infers miRNA binding and Degradome-seq read counts from the mRNA sequence (Kanuparthi et al., 2025). However, when applied to experimental human and mouse data across 39 cell lines and tissues, the model does not take the miRNA sequence as input and predicts only miRNA target sites that are highly expressed in the specific cell lines and tissues under study, which limits its ability to generalize to novel miRNA target sites. Overall, existing methods mainly suffer from two fundamental limitations: reliance on 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 attention mechanisms (Ji et al., 2021; Dalla-Torre et al., 2025; Avsec et al., 2021; Nguyen et al., 2024; Patel et al., 2025; Wang et al., 2024; Akiyama & Sakakibara, 2022). In particular, transformer-based RNA language models are highly effective at capturing miRNA-mRNA interactions from raw sequences. Mimosa leverages two independent encoders - one encodes mRNA sequence and the other encodes miRNA sequence, and discovers local target sites by Smith-Waterman local alignment. This approach is a unique combination of both advanced deep-learning and classic alignment algorithm (Bi et al., 2025). RNAErnie, a transformer-based RNA language model pretrained on multi-level motif-aware masking strategies, outperforms both heuristic methods and other transformer-based RNA language models on predicting miRNA-mRNA interactions (Wang et al., 2024).

In this study, we introduce **MiRformer**, a hybrid convolution–transformer framework that jointly predicts sequence-level miRNA–mRNA interactions and nucleotide-level miRNA binding sites (i.e., seed regions), as well as in-silico synthesizes target-specific miRNA given any mRNA. MiRformer employs a dual-transformer-encoder architecture that processes miRNA and mRNA sequences independently. To efficiently scale to long mRNA sequences, we adopt sliding-window local self-attention with reduced computational complexity (Beltagy et al., 2020). To explicitly model nucleotide-level miRNA–mRNA interactions, we further propose a sliding-window cross-attention mechanism with Log-Sum-Exponential (LSE) pooling to preserve strong local alignment signals and improve interpretability. Across benchmarks on TargetScan and experimental Degradome-seq datasets, MiRformer achieves state-of-the-art performance on binding prediction, seed-region localization, and cleavage-site identification. Extending MiRformer, we repurpose MiRformer's components and train a miRNA decoder that, given an mRNA as input, autoregressively generates a miRNA specific to that mRNA. Across 5712 target mRNAs, MiRformer achieves a 99.30% seed match rate for all predicted miRNAs, and close to 50% of them are canonical 8-mer seed matches.

2 METHODS

2.1 MIRFORMER MODEL

MiRformer is a dual-transformer framework designed to model miRNA–mRNA interactions from paired nucleotide sequences (Fig. A1a). Both miRNA and mRNA sequences are tokenized at the nucleotide level and passed through convolutional layers to enhance local sequence continuity, following prior observations that convolution-augmented embeddings improve token coherence in RNA modeling (Zhang et al., 2024). MiRformer employs two transformer encoders trained with asymmetric attention mechanisms to reflect the distinct roles of miRNA and mRNA. The miRNA encoder uses full self-attention to capture global dependencies within the short miRNA sequence, while the mRNA encoder adopts sliding-window self-attention to efficiently scale to long mRNA contexts and model local binding regions (Fig. A1b), following the Longformer design (Beltagy et al., 2020).

To explicitly model nucleotide-level miRNA–mRNA interactions, MiRformer introduces a sliding-window cross-attention mechanism that allows miRNA tokens to attend to local mRNA segments. Overlapping cross-attention windows are merged using Log-Sum-Exponential (LSE) pooling, which preserves strong local alignment signals while aggregating information across windows. This design yields interpretable attention maps that highlight biologically meaningful seed regions without incurring the quadratic cost of global attention.

To extend on MiRformer's ability on molecular design, we designed MiRformer to generate miRNAs given any target mRNA as input. The generative architecture comprises four main modules: an mRNA tokenizer, a miRNA tokenizer, an mRNA encoder, and a miRNA decoder (Fig. A1d). It reuses the pre-trained mRNA encoder from MiRformer and trains a miRNA decoder. In contrast to MiRformer, the miRNA tokenizer in MiRformer omits convolutional layers to avoid accessing information from future miRNA bases. A comparison of MiRformer variants with and without convolution in the miRNA tokenizer shows that removing convolution substantially improves per-base prediction accuracy (Fig. A6). Full architectural details and mathematical formulations are provided in Appendix A.1

2.2 PREDICTION TASKS

The three prediction tasks are illustrated in Fig. A1c and described here.

Target prediction We curated a balanced dataset of 560k human and mouse miRNA-mRNA pairs from TargetScan and splited into train/validation/test sets (90/8/2) with more details in Appendix A.2.1. Given a pair of miRNA and mRNA, 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 into a linear classifier, which outputs a sequence-level sigmoid probability.

Seed-region recognition Given a pair of miRNA and target mRNA, MiRformer was trained to predict the start and end positions of the seed region within the mRNA sequence. The nucleotide-level embedding is fed into 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.

Degradation event prediction We mapped human degradome-seq data to 3'UTR transcript coordinates using GENCODE v19 (GRCh37) annotations, ultimately sampling 210k 500-nt windows containing cleavage sites for a 90/8/2 training, validation, and test split. Further details are in Appendix A.2.2. Given a pair of miRNA and target mRNA, MiRformer was trained to predict cleavage sites coordinates within 500-nt windows of 3'UTR mRNAs.

3 RESULTS

3.1 HYBRID EMBEDDING IMPROVES MODEL PERFORMANCE AND INTERPRETABILITY

As a proof-of-concept, we evaluated the effect of hybrid tokenization on short 30-nt mRNAs (Fig. A2). Models with convolution outperformed ablated variants lacking convolution (Fig. A2a). Analysis of cross-attention maps showed that ablated models primarily attended to seed boundaries,

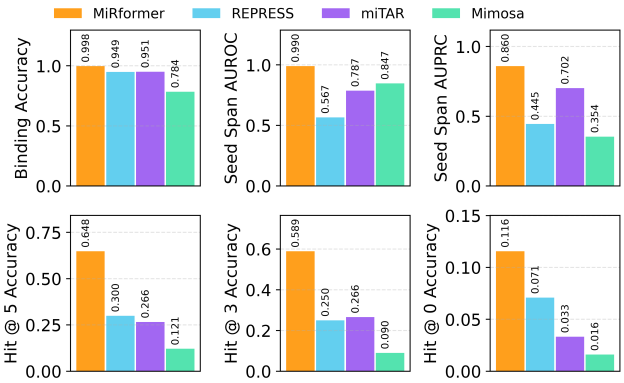

Figure 1: **Benchmark on seed region and cleavage site prediction.** Binding Accuracy evaluates classification of miRNA-mRNA interactions. Seed Span AUROC and AUPRC evaluate nucleotide-level prediction for inside the seed region. Hit @ k Accuracy measure accuracies of predicting the cleavage site within k-nt distance to the true cleavage site in the target mRNA sequence.

whereas MiRformer with convolution layers exhibited strong attention over both the boundaries and interior of the seed region (Fig. A2b–c). To further assess interpretability at single-nucleotide resolution, we performed in-silico mutagenesis (ISM; Section A.5). Changes in target prediction scores and attention weights were highest within seed regions (Fig. A3a–c), indicating that MiRformer identifies key binding nucleotides. Additionally, inserting artificial seed regions into negative miRNA–mRNA pairs led to significantly increased target prediction scores (Fig. A3d), confirming that the model relies on true seed features. Overall, convolutional layers enable effective local feature extraction that improves both predictive performance 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. A4). In conclusion, we showed that LSE improved all metrics compared to mean pooling and performed on-par with global cross attention while achieving better seed recognition accuracy (0.925 vs 0.919; Fig.A4c). 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 assess whether MiRformer captures seed regions in long (500-nt) mRNA contexts, we analyzed sliding-window cross-attention patterns and performed in-silico mutagenesis (ISM) on representative examples (Fig. A5). Both attention scores and ISM signals consistently concentrated on seed regions, indicating that MiRformer relies on biologically meaningful nucleotides. We evaluated two attention normalization schemes: normalization by mRNA highlights which mRNA positions are most attended by the miRNA, and normalization by miRNA emphasizes which miRNA nucleotides drive binding(Fig.A5a,b). Both reveal clear seed-specific patterns, with sharper boundary localization under miRNA-normalized attention. In contrast, mean pooling produces substantially noisier attention map(Fig.A5c), and although global attention achieves high predictive performance, its attention patterns are not interpretable(Fig.A5d). Quantitatively, sliding-window cross-attention with LSE pooling yields a substantially larger separation between seed and non-seed attention scores than mean pooling, confirming that LSE preserves strong local alignment signals (Fig. A5a-c). Together, these results demonstrate that MiRformer achieves an effective trade-off between interpretability and predictive performance.

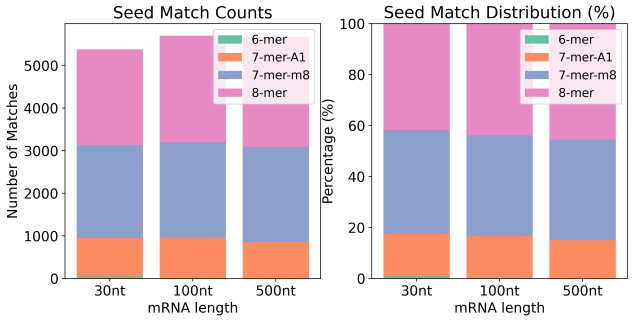

Figure 2: **Distribution of canonical seed match types in MiRformer-synthesized miRNAs.** Seed match categories include 6mer, 7mer-a1, 7mer-m8, and 8mer, shown as the counts of each category and their percentages relative to the total number of seed matches.

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

We evaluated MiRformer against three state-of-the-art models (REPRESS, miTAR and Mimosa) for 3 tasks - target prediction, seed region recognition and experimental degradation event prediction (Section 2.2) and measured all models performance by 6 metrics (Fig. 1). For the first two tasks, we used the TargetScan Human and Mouse (version 7.2) as labels (AppendixA.2.1). All methods were tested on 500-nt mRNA windows of 3'UTR. MiRformer outperformed other models on all 6 metrics (Fig. 1). 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 0.323 (Fig. 1 bottom).

4 MiRformer synthesizes target-specific miRNA candidates by re-purposing MiRformer

Extending on MiRformer, we evaluated the model's capacity to in-silico synthesize miRNA molecules. We modified MiRformer by repurposing MiRformer's components to generate miRNA for any given target mRNAs. MiRformer fine-tunes mRNA encoder with a sliding-window self-attention mechanism and trains a dedicated miRNA decoder to autoregressively synthesize miRNA sequences nucleotide by nucleotide (Fig.A1d). We assessed MiRformer on three mRNA lengths—30 nt, 100 nt, and 500 nt—and observed that per-base miRNA prediction accuracy increases as the length of the input mRNA decreases (Fig.A7). We hypothesized that longer mRNAs introduce more noise and may contain additional, unlabeled seed matches. To examine this, we aligned miRNA seed regions—positions 2–8 from the 5' to 3' end of the miRNA—to the corresponding target mRNAs and quantified four canonical seed match types: 6mer, 7mer-a1, 7mer-m8, and 8mer. We found that the majority of MiRformer-generated miRNAs harbor valid seed regions that are complementary to their target mRNAs, achieving 99.30% seed match coverage across 5,712 tested 500-nt mRNA targets, with 45.49% of these matches being 8mers (Fig.2). Raw seed match counts and match rates for each seed type are reported in Table2. Together, these results demonstrate MiRformer's ability to synthesize biologically plausible miRNAs that contain complementary seed regions to their target mRNAs.

5 Discussion

Characterizing the sequence determinants of miRNA–mRNA interactions remains a long-standing challenge. Traditional methods such as TargetScan rely on handcrafted features and heuristic rules, while existing deep learning approaches often struggle to model nucleotide-resolution interactions in long mRNA contexts. Transformer-based models offer a promising alternative by capturing both local and global dependencies through attention mechanisms.

In this work, we introduced MiRformer, a unified generative framework that models miRNA and mRNA interactions and synthesizes mRNA-conditioned miRNA using sliding-window cross-attention with LSE pooling. This design scales efficiently to long mRNA sequences while producing

interpretable attention patterns. Across benchmarks, MiRformer achieves accurate and interpretable predictions for binding, seed localization, and cleavage-site identification. Building on MiRformer, we finetuned MiRformer's modules to in-silico synthesize miRNA for any given target mRNA. Almost all MiRformer-synthesized miRNAs contain valid seed match given 500nt mRNA targets.

A current limitation is that MiRformer is primarily trained on canonical seed regions derived from TargetScan, which may restrict generalization to non-canonical interactions. Incorporating larger collections of experimentally validated datasets (Huang et al., 2022; Li et al., 2014) and leveraging cross-species multiple sequence alignments, as explored in recent genome language models (Benegas et al., 2025), may improve detection of conserved and non-canonical target sites.

Looking ahead, MiRformer could jointly learn RNA structures from raw sequences, possibly through multimodal transformer, without having to pre-compute heuristic features such as minimum energy. It can be extended to incorporate single-cell expression profiles (Hücker et al., 2021). Pretrained foundational RNA language models across diverse transcriptomes may enable zero-shot regulatory inference for novel miRNAs and disease-associated variants (Benegas et al., 2025). We view MiRformer and MiRformer as a step toward more general, data-driven modeling of post-transcriptional gene regulation.

REFERENCES

Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. Predicting effective microrna target sites in mammalian mrnas. *elife*, 4:e05005, 2015. URL https://doi.org/10.7554/eLife.05005.

Manato Akiyama and Yasubumi Sakakibara. Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. *NAR Genomics and Bioinformatics*, 4 (1):lqac012, 2022. ISSN 2631-9268. doi: 10.1093/nargab/lqac012. URL https://doi.org/10.1093/nargab/lqac012.

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. *Nature Methods*, 18 (10):1196–1203, 2021. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-021-01252-x.

David P Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297, 2004. URL https://doi.org/10.1016/S0092-8674(04)00045-5.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. (arXiv:2004.05150), 2020. doi: 10.48550/arXiv.2004.05150. URL http://arxiv.org/abs/2004.05150.

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. *Nature Biotechnology*, 2025. ISSN 1546-1696. doi: 10.1038/s41587-024-02511-w. URL https://doi.org/10.1038/s41587-024-02511-w.

Yue Bi, Fuyi Li, Cong Wang, Tong Pan, Chen Davidovich, Geoffrey I. Webb, and Jiangning Song. Advancing microRNA target site prediction with transformer and base-pairing patterns. *Nucleic Acids Research*, 2025. URL https://dx.doi.org/10.1093/nar/gkae782.

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. *Nature Methods*, 22(2):287–297, 2025. ISSN 1548-7105. URL https://doi.org/10.1038/s41592-024-02523-z.

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. URL https://doi.org/10.1186/s12859-021-04026-6.

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. URL https://doi.org/10.1002/hep.26302DigitalObjectIdentifier(DOI).

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. URL http://www.genome.org/cgi/doi/10.1101/gr.135350.111.

Hsi-Yuan Huang, Yang-Chi-Dung Lin, Shidong Cui, Yixian Huang, Yun Tang, Jiatong Xu, Jiayang Bao, Yulin Li, Jia Wen, Huali Zuo, Weijuan Wang, Jing Li, Jie Ni, Yini Ruan, Liping Li, Yidan Chen, Yueyang Xie, Zihao Zhu, Xiaoxuan Cai, Xinyi Chen, Lantian Yao, Yigang Chen, Yijun Luo, Shupeng LuXu, Mengqi Luo, Chih-Min Chiu, Kun Ma, Lizhe Zhu, Gui-Juan Cheng, Chen Bai, Ying-Chih Chiang, Liping Wang, Fengxiang Wei, Tzong-Yi Lee, and Hsien-Da Huang. miRTarBase update 2022: an informative resource for experimentally validated miRNA–target interactions. *Nucleic Acids Research*, 50:D222–D230, 2022. ISSN 0305-1048. doi: 10.1093/nar/gkab1079. URL https://doi.org/10.1093/nar/gkab1079.

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. URL https://doi.org/10.1038/s41467-021-24611-w.

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. URL https://doi.org/10.1093/bioinformatics/btab083.

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*, pp. 2025–05, 2025. URL https://doi.org/10.1101/2025.05.15.654105.

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. URL https://doi.org/10.1093/nar/gkt1248.

Shun Liu, Jun-Hao Li, Jie Wu, Ke-Ren Zhou, Hui Zhou, Jian-Hua Yang, and Liang-Hu Qu. StarScan: a web server for scanning small RNA targets from degradome sequencing data. *Nucleic Acids Research*, 43:W480–W486, 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv524. URL https://doi.org/10.1093/nar/gkv524.

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. *Science*, 386(6723):eado9336, 2024. doi: 10.1126/science.ado9336. URL https://www.science.org/doi/10.1126/science.ado9336.

Sawan Patel, Fred Zhangzhi Peng, Keith Fraser, Adam David Friedman, Pranam Chatterjee, and Sherwood Yao. EvoFlow-RNA: Generating and representing non-coding RNA with a language model. In *ICLR 2025 Workshop on AI for Nucleic Acids (AI4NA)*, 2025. URL https://openreview.net/forum?id=oDTBI7z52K.

Albert Pla, Xiangfu Zhong, and Simon Rayner. miRAW: A deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts. *PLOS Computational Biology*, 14(7):e1006185, 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006185. URL https://doi.org/10.1371/journal.pcbi.1006185.

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. URL https://doi.org/10.1002/hep.26712.

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. *Nature Machine Intelligence*, 6(5):548–557, 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00836-4. URL https://www.nature.com/articles/s42256-024-00836-4.

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. URL https://doi.org/10.1093/bioinformatics/bty424.

Zaixi Zhang, Linlin Chao, R Jin, Yikun Zhang, Guowei Zhou, Yujie Yang, Yukang Yang, Kaixuan Huang, Qirong Yang, and Ziyao Xu. Rnagenesis: foundation model for enhanced rna sequence generation and structural insights. *bioRxiv*, pp. 2024–12, 2024. URL https://doi.org/10.1101/2024.12.30.630826.

A Appendix

A.1 MiRformer framework details

Convolution layers 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. A1a). This design is inspired by RNAGenesis (Zhang et al., 2024), which demonstrated that convolution-augmented embeddings improve token coherence.

A.1.1 Dual-transformer-encoder architecture with distinct self-attention mechanisms

miRFormer uses two dedicated transformer encoders that are jointly trained end-to-end (Fig. A1a) with distinct attention mechanisms: **full self-attention** and **sliding-window attention** for miRNA and mRNA, respectively (Fig. A1b). 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 (Beltagy et al., 2020), 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.

A.1.2 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 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. A1b). 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. Finally, attention is normalized 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 we explored (Fig. A5).

A.2 Datasets

A.2.1 TargetScan dataset

From TargetScan v7.2, we curated a dataset of 560k human and mouse miRNA-mRNA pairs, comprising 280k positive pairs containing miRNA target sites and 280k negative pairs lacking miRNA target sites. Positive samples were labeled with 1 and the seed start and seed end positions within each mRNA. Negative samples were labeled as 0 and the seed positions were set to -1 and were excluded from model training. Before training, the data were divided into separate training, validation, and test sets. The split ratio was 0.9 for training, 0.08 for validation, and 0.02 for testing. Only training and validation sets were seen during model training. We considered window lengths of 30 nt and 500 nt to evaluate both short- and long-context modeling.

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 from the full 3'UTR sequence. If the 3'UTR length was shorter than the window length, we retained the full sequence. To reduce degree/composition bias introduced by global shuffling, we constructed negative samples in a miRNA-stratified, degree-preserving manner. For each miRNA, we sampled an equal number of negative mRNAs, where a negative mRNA is defined as an mRNA not predicted as a target of that miRNA by TargetScan. This design preserves the miRNA marginal frequency between positive and negative pairs and reduces the possibility that the model memorizes heterogeneous miRNA interaction degrees. Given a negative (miRNA, mRNA) pair, we then extracted seed-absent mRNA segments from the full 3'UTR sequence. If the 3'UTR length was shorter than the window length, we retained the full sequence. Otherwise, we sampled up to N windows of the specified length such that no window contains a $\geq$ 6-nt consecutive complementary seed match to the paired miRNA. We set $N = 1$ for 30-nt windows and $N = 8$ for 500-nt windows, with $N \leq L - w + 1$ where L is the 3'UTR length and w is the window length. Overall, we merged human and mouse miRNA–mRNA pairs to build a dataset of 560k samples, consisting of 280k positive pairs with confirmed miRNA seed regions and 280k negative pairs without seed regions.

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.

A.2.2 Degradome-seq dataset

Human degradome-seq data, including mRNA–miRNA pairs and experimental genomic cleavage site coordinates, were obtained from StarBase and StarScan (Li et al., 2014; Liu et al., 2015). The 3'UTR sequences and gene coordinates were extracted from GENCODE v19 annotations (Harrow et al., 2012) for reference genome assembly GRCh37 (hg19). Cleavage site genomic coordinates were subsequently converted to 3'UTR transcript coordinates using the same GENCODE v19 reference. The length of the longest sequence reached 25k nucleotides. Next, we constructed a 500-nt dataset by sampling 500-nt windows containing the cleavage sites from 3'UTR. We collected 210k miRNA-mRNA pairs with observed cleavage sites. The 500-nt datasets were partitioned into training, validation, and test sets using a split ratio of 0.9 for training, 0.08 for validation, and 0.02 for testing. Only training and validation sets were used for model training.

A.3 OBJECTIVE FUNCTIONS

MiRformer minimizes a combined loss function for **Target prediction** and **Seed-region recognition** 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 $\sum_{i=1}^L \frac{w_i\,\mathcal{L}_i}{\sum_{i=1}^L w_i}$.

A.4 HYPERPARAMETER TUNING

We optimized the model's hyperparameters to identify the configuration that yielded the best overall performance—specifically, the set of values that minimized training loss and, on the validation set, achieved the highest binding accuracy and F1 score for overlap between the predicted seed match and the TargetScan seed match (Table 1).

A.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 $f \in \{1, 2\}$.

A.6 BASELINE METHODS

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 (Kanuparthi et al., 2025) - average pooled both the prediction and labels within the predefined window lengths for each cell line and divided the output into windows. For Mimosa, we slided a 40-nt window across each mRNA, scoring each window with the pretrained Mimosa Transformer, and classify the miRNA-mRNA as interactive if the best window's probability exceeds a threshold. We predicted seed start and end by inferring the paired region within the best-scoring window (via the same seed-alignment pairing map used by Mimosa) and aggregated per-nucleotide seed-span scores across windows. In a similar manner, we predicted cleavage sites by identifying the cleavage position within the best-scoring window based on seed alignment and mapping the inferred site back to the original mRNA sequence.

Table 1: Optimal hyperparameters for MiRformer model.

Parameter	30-nt	500-nt
Embedding dimension	256	1024
Feed-forward dimension	512	4096
Number of encoder layers	2	4
Number of attention heads	2	8
Learning rate	1×10^{-4}	1×10^{-4}–1×10^{-5}
Weight decay	0.1	0.1
Batch size	256	256
Maximum sequence length	30	500

Table 2: Comparison of predicted miRNA seed match types across different mRNA lengths

Metric	**30nt**	**100nt**	**500nt**
Raw Counts			
Valid 6-mers	48	17	8
Valid 7-mers-A1	902	939	844
Valid 7-mers-m8	2,180	2,244	2,240
Valid 8-mers	2,249	2,495	2,580
Total Functional Match Rate	**94.17%**	**99.70%**	**99.30%**
Distribution (% of Total Matches)			
Percentage of 6-mers	0.89%	0.30%	0.14%
Percentage of 7-mers-A1	16.77%	16.49%	14.88%
Percentage of 7-mers-m8	40.53%	39.40%	39.49%
Percentage of 8-mers	41.81%	43.81%	45.49%

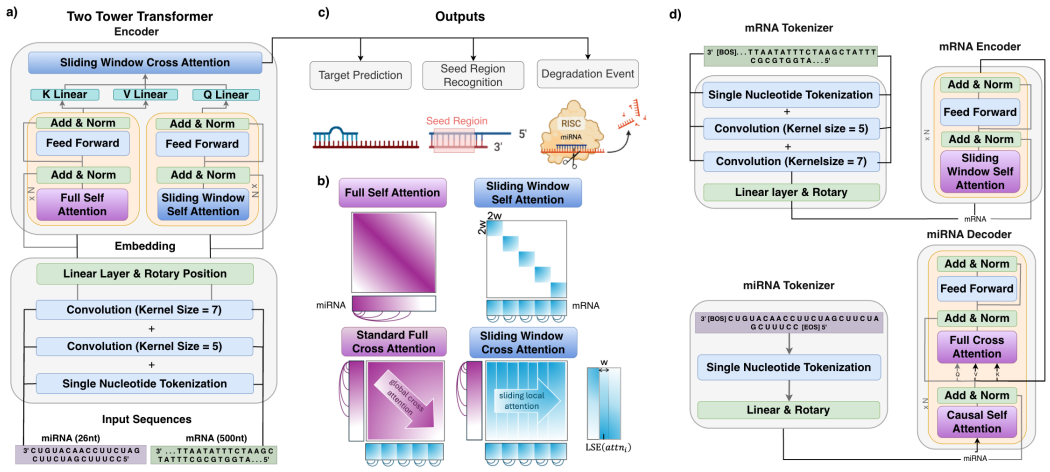

Figure A1: **MiRformer model overview** a) MiRformer leverages convolutional tokenization layers to extract seqeunce features from miRNA and mRNA, and encodes the features via a Dual-Transformer-Encoder Architecture (DTEA). The resulting embeddings are then fed to two dedicated transformer encoders for 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, where $w = 20$ or a chosen overlapping width. While full cross attention attends to the entire miRNA and mRNA embeddings, sliding-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. d) MiRformer leverages mRNA Tokenizer and finetunes MiRformer's mRNA Encoder. It trains a miRNA decoder to synthesize miRNA given a target mRNA as prompt.

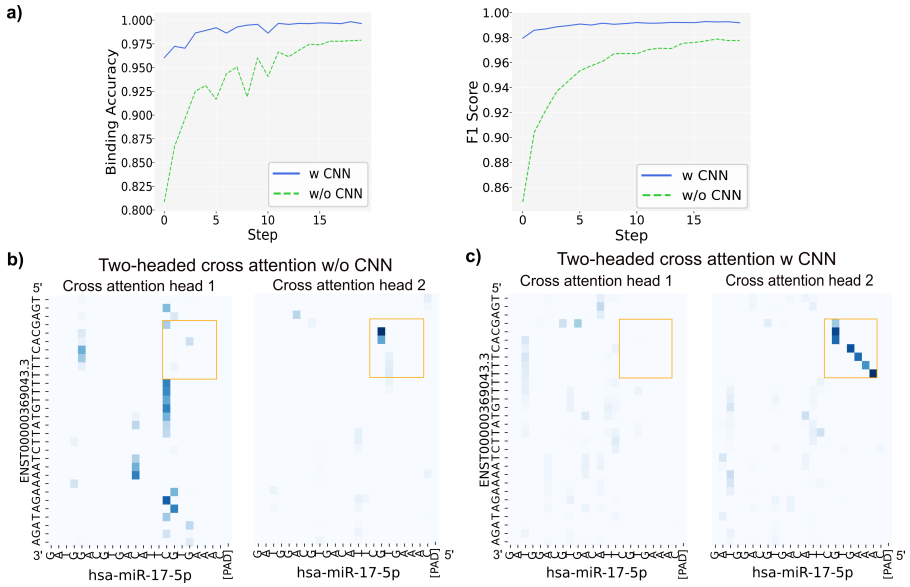

Figure A2: **Proof-of-concept analysis on 30-nt short mRNA sequences.** a) Model performance comparison between with and without convolutional feature extraction. b) Heatmap produced by model trained without convolution feature embeddings. c) Heatmap produced by model with convolution feature embeddings. 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.

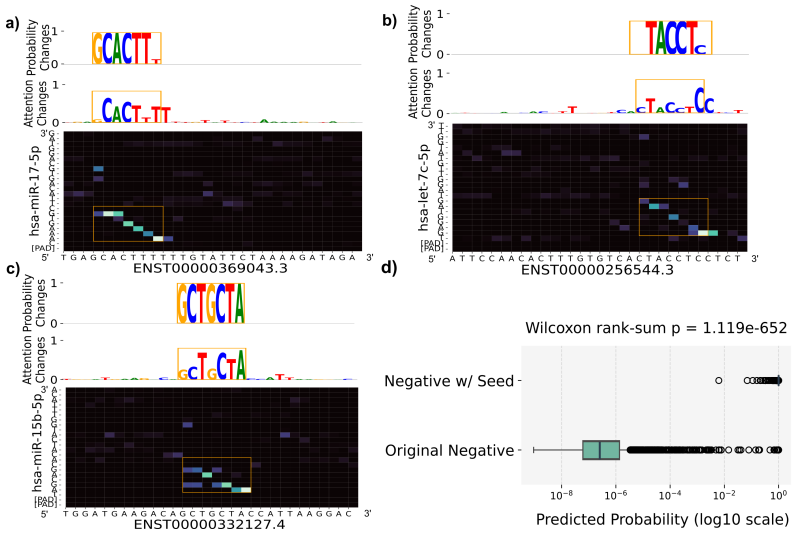

Figure A3: **In-Silico Mutagenesis (ISM) analysis of 30-nt mRNA sequences.** (a-c) ISM and attention visualization. We perturbed every base and recorded the change in target prediction probability and the changes in cross attention matrix at each base. For each of the 3 examples, the top two tracks display 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.

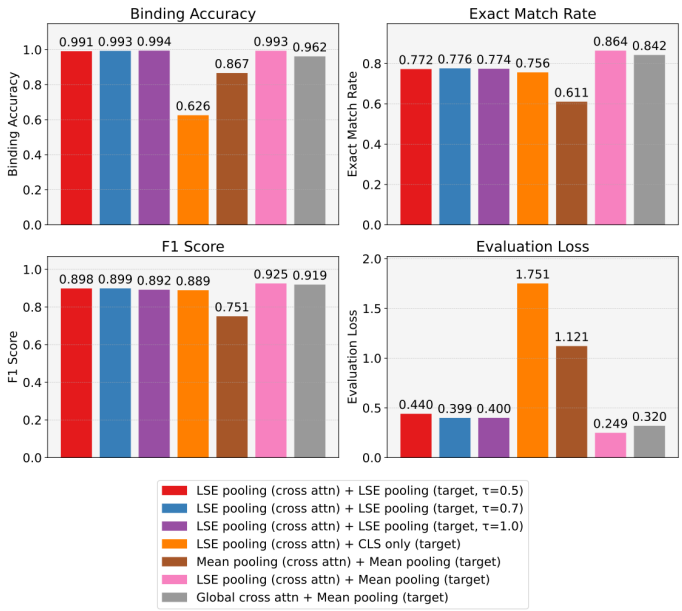

Figure A4: **Pooling methods comparison.** The pooling method is followed by the model layer to which it is applied (for example, LSE pooling (cross attn) indicates that LogSumExp pooling is used in the cross-attention layer, while Mean pooling (target) indicates that mean pooling is used in the target prediction head). τ denotes the temperature parameter used in LogSumExp pooling. a) Accuracy of the predicted miRNA–mRNA interactions. b) Accuracy of the predicted seed start and seed end positions. c) F1 score quantifying the overlap between predicted seed regions and the true seed regions. d) Loss for each method, evaluated on an independent validation set for both target and seed predictions.

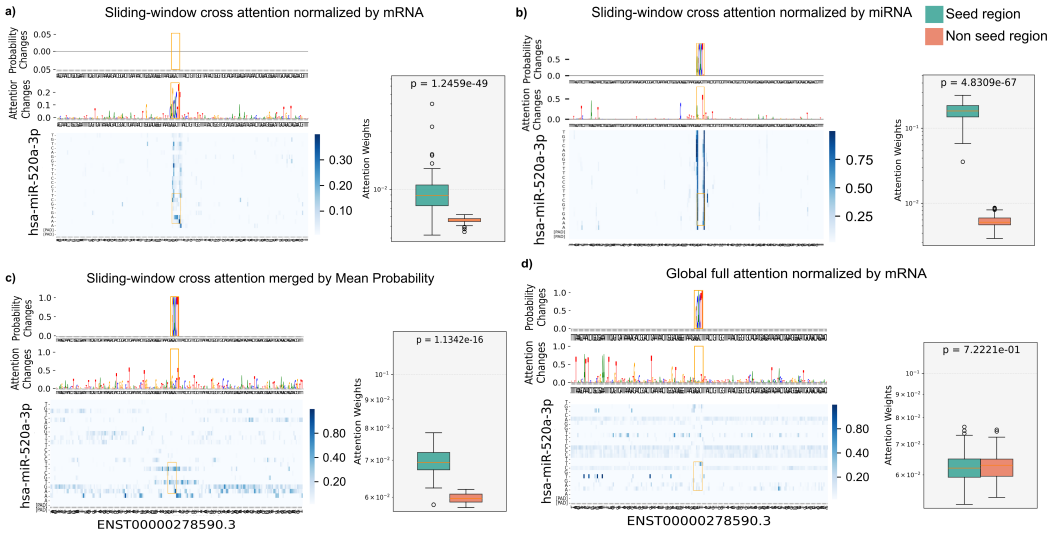

Figure A5: **Model interpretability through cross attention heatmaps on a 500-nt mRNA sequence.** For each panel, top track displays the change in probability at the single nucleotide resolution after ISM; middle track shows the change in attention after ISM; bottom track displays cross attention between 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.

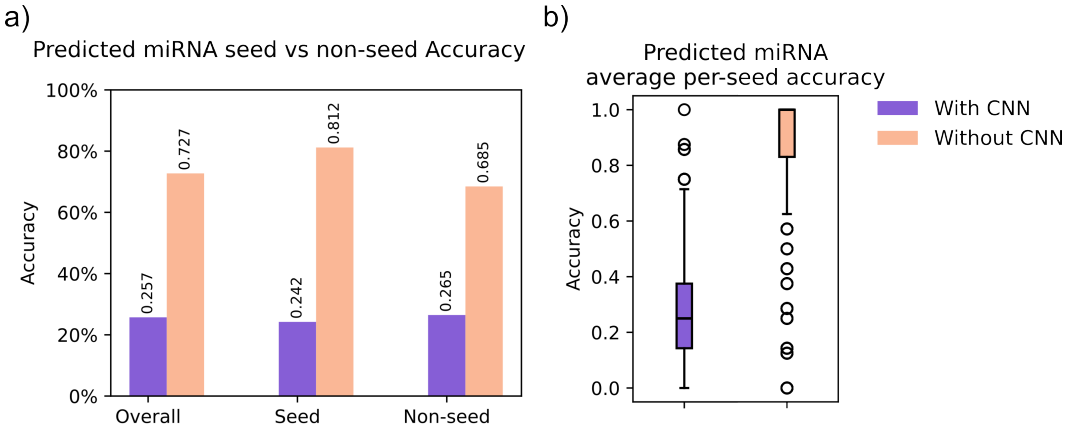

Figure A6: Comparison of generated miRNA accuracy between MiRformer with and without convolution. a) Comparison of predicted miRNA per-base accuracy overall regions, seed and non-seed regions between MiRformer with and without CNN. b) Comparison of distribution of predicted miRNA per-base accuracy in the seed regions between MiRformer with and without convolution.

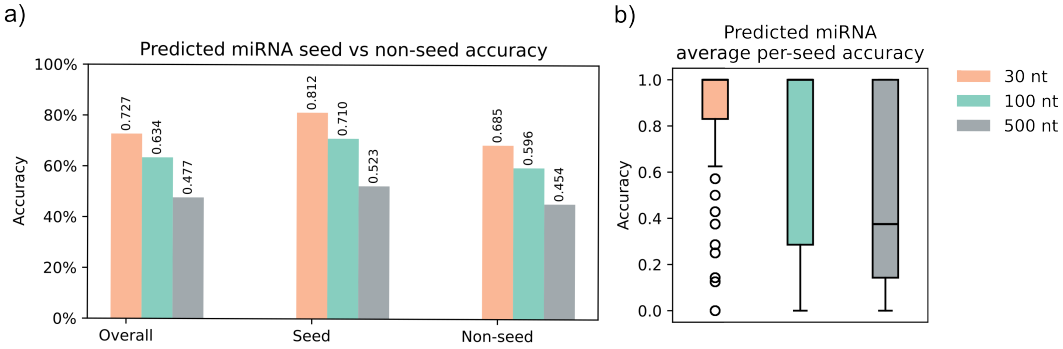

Figure A7: Comparison of generated miRNA accuracy between 30nt, 100nt, 500nt mRNA lengths. a) Comparison of predicted miRNA per-base accuracy overall regions, seed and non-seed regions between 30nt, 100nt and 500nt mRNAs. b) Comparison of distribution of predicted miRNA per-base accuracy in the seed regions between 30nt, 100nt and 500nt mRNAs.