MMLmiRLocNet: miRNA Subcellular Localization Prediction based on Multi-view Multi-label Learning for Drug Design

Tao Bai, Junxi Xie, Yumeng Liu*, Bin Liu*

***Abstract*—Identifying subcellular localization of microRNAs (miRNAs) is essential for comprehensive understanding of cellular function and has significant implications for drug design. In the past, several computational methods for miRNA subcellular localization is being used for uncovering multiple facets of RNA function to facilitate the biological applications. Unfortunately, most existing classification methods rely on a single sequence-based view, making the effective fusion of data from multiple heterogeneous networks a primary challenge. Inspired by multi-view multi-label learning strategy, we propose a computational method, named MMLmiRLocNet, for predicting the subcellular localizations of miRNAs. The MMLmiRLocNet predictor extracts multi-perspective sequence representations by analyzing lexical, syntactic, and semantic aspects of biological sequences. Specifically, it integrates lexical attributes derived from k-mer physicochemical profiles, syntactic characteristics obtained via word2vec embeddings, and semantic representations generated by pre-trained feature embeddings. Finally, module for extracting multi-view consensus-level features and specific-level features was constructed to capture consensus and specific features from various perspectives. The full connection networks are utilized as the output module to predict the miRNA subcellular localization. Experimental results suggest that MMLmiRLocNet outperforms existing methods in terms of F1, subACC, and Accuracy, and achieves best performance with the help of multi-view consensus features and specific features extract network. The web server of MMLmiRLocNet has been established, which can be accessed at http://bliulab.net/MMLmiRLocNet.**

***Index Terms*—subcellular localization of miRNA; multi-view feature learning; deep learning; multi-label learning**

I. Introduction

MicroRNAs (miRNAs), also known as short non-coding RNAs, are fundamental to various cellular processes in animals and plants, such as development, digestion, proliferation, and differentiation [1-4]. They are integral to post-transcriptional gene regulation [5, 6]. The subcellular localization of miRNAs is crucial as it influences their interaction with proteins and other RNAs, thereby affecting their function [7, 8]. These interactions, whether direct or indirect, influence miRNA function [9, 10]. For instance, mitochondrial miRNAs are involved in mitochondrial metabolism [11]. Additionally, exosomes and microvesicles act as vectors for cell-to-cell communication and potential biomarkers for diseases [12-14]. MiRNA is also pivotal in drug design [15]. Recent research underscores the importance of miRNA subcellular localization in understanding diseases [16] such as Alzheimer's disease [17] and cancer [18], suggesting that manipulating miRNA subcellular localization may offer novel therapeutic strategies for currently underserved diseases [19, 20]. However, compared to the other ncRNAs, the number of studies related to the subcellular localization of miRNAs remains limited due to single-perspective sequence representations, lack of ontology, and few functional annotations [21-23]. Therefore, there is an urgent need to design a comprehensive computational tool for miRNA subcellular localization.

Recently, several computational approaches have been established to analyze the subcellular localization of miRNAs. Methods such as MiRLocator [24], MirLocPredictor [25], L2S-MirLoc [26], miRNALoc [27], MulStack [28], and iLoc-miRNA [29] rely on sequence-based feature, including intrinsic composition and physicochemical properties, such as one-hot encoding [30-32], k-mer [33], and Electron-ion Interaction Pseudo Potentials(EIIP) [34], etc. However, these methods lack sufficient sequence feature similarity and effective structural information. On the other hand, methods like MiRGOFS [35] and MiRLoc [36] focus on miRNA functional similarity networks but lack sufficient sequence information. DAmiRLocGNet [37] is the only method that effectively combines sequence representation with functional similarity networks to predict miRNA subcellular localization. Recent studies have demonstrated that the physicochemical properties of miRNA sequence information and miRNA functional networks contribute to miRNA subcellular location prediction

This work was supported by the National Natural Science Foundation of China (No. U22A2039, 62325202 and No. 62302316), Shenzhen Science and Technology Program (Grant No.RCBS20221008093227027) and Natural Science Foundation of Top Talent of SZTU (Grant No. GDRC202319) *(Corresponding authors: Yumeng Liu and Bin Liu)*.

Tao Bai is with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China; and the School of Mathematics and Computer Science, Yan'an University, Yan'an, 716000, China(e-mail: tbai@bliulab.net).

Junxi Xie and Yumeng Liu are with the College of Big Data and Internet, Shenzhen Technology University, Shenzhen 518118, Guangdong, China (e-mail: 2310413032@stumail.sztu.edu.cn and liuyumeng@sztu.edu.cn).

Bin Liu is with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China, and Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, Shenzhen 518172, China (e-mail: bliu@bliulab.net).

[38, 39]. Nonetheless, there is still a pressing need for research to improve prediction accuracy by integrating these features more effectively.

To the best of our knowledge, identifying RNA subcellular localization should be addressed as a multi-label classification problem [21, 40-44]. In this context, Multi-label learning is a fundamental framework that assigns multiple labels to a single instance [45-47]. Multi-view data, which represent the same entity from various perspectives, offer a more comprehensive depiction of multi-label classification issues compared to single-view data [48]. Studies have demonstrated that diverse feature representations of biological sequences from multiple perspectives enhance classification performance [49-53]. Therefore, employing a multi- perspective heterogeneous data feature fusion network can improve the prediction performance.

In this paper, inspired by multi-view multi-label learning approach [45, 54, 55], we present our methodology, MMLmiRLocNet, as illustrated schematically in Fig. 1. We make an attempt to use consensus features and specific features extract networks to predict multiple subcellular localizations of miRNA based on multi-view multi-label learning framework. The MMLmiRLocNet framework captures both consensus-level features and specific-level features from multiple views of biomedical sequences, encompassing lexical, syntactic, and semantic representations. Specifically, it includes lexical properties derived from k-mer physicochemical features, syntactic features obtained from word2vec embeddings, and semantic representations generated by pre-trained BERT algorithm, respectively. Experimental results demonstrate that the MMLmiRLocNet outperforms previous predictor in terms of F1, subACC, and Accuracy, achieving high performance through the integration of multi-view consensus features and specific feature extraction networks. Additionally, the web server and source code of MMLmiRLocNet can be freely visited at http://bliulab.net/MMLmiRLocNet.

II. Methods

A. Motivation

Inspired by the similarity between natural language and biological sequences [56-59], our motivation is to enhance the feature representation by extracting diverse perspectives—lexical, syntactic, and semantic—form biological sequences. This approach motivation is depicted in Fig. 1. From the lexical perspective, it focuses on word properties, frequencies, distributions, and morphological features to capture essential lexical characteristics, thereby significantly improving sequence comprehension. Conversely, the syntax perspective explores the structural relationships between words, parses sentences, and identifies grammatical structures, which reveal underlying syntactic patterns that shape sequences. Lastly, the semantic perspective delves into word meanings, contextual relationships, and semantic roles to uncover deeper semantic representations, offering insights into sequence content and intent. The role of features from multiple views in label prediction varies. Considering the contribution of different views, focusing on the private specific features of lexical, semantic, consensus features of syntactic, and helps in determining whether the corresponding label is microvesicles.

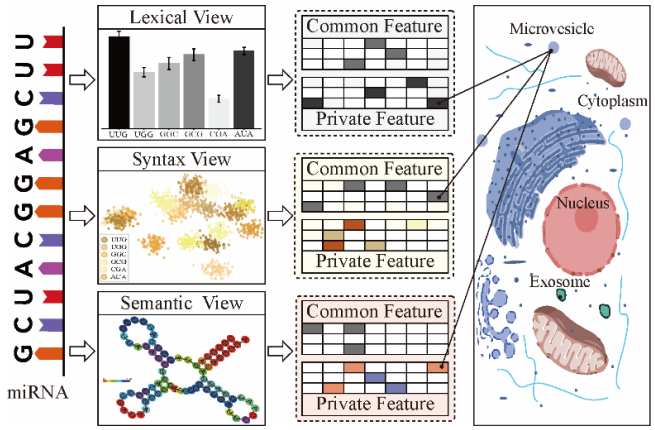

Fig 1. The motivation of method. The Lexical View focuses on analyzing individual words and their properties. The Syntax View investigates structural relationships between words. The Semantic View delves into meanings and contextual relationships.

B. Dataset

1) Dataset collection: The initial crucial step in developing a reliable predictor involves establishing a robust benchmark dataset. To accomplish this, miRNA subcellular localization data were extracted from the RNALocate v2.0 database [60] (https://www.rna-society.org/rnalocate/). This database consolidates over 210,000 experimentally validated RNA-associated subcellular localization records, covering more than 110,000 RNAs across 171 subcellular localizations and encompasses data from 104 species. The dataset was prepared according to the following strategy:

a) We retrieved 12,781 entries of Homo sapiens miRNA-associated subcellular localization from the RNALocate v2.0 database. To mitigate the redundancy, miRNAs with multiple entries, those miRNAs with identical gene symbols and subcellular localizations were merged.

b) We deleted the miRNAs lacking sequence information in miRbase [61]. To further reduce data redundancy, we utilized the CD-HIT-EST [62] with the following parameters: *-c 0.80 -n 5*.

c) Given the limited number of miRNA entries for specific subcellular locations, we retained only those with more than 80 occurrences.

Finally, we obtained 538 miRNAs, including nucleus, exosome, cytoplasm, and microvesicle. The distribution of the miRNA subcellular localization dataset is summarised in Fig. 2.

As illustrated in Fig. 2(a), an upset plot depicts the distribution of miRNA sequences across various intersection groups, arranged by compartment numbers. The upper bar plot displays the count of miRNAs in each intersection group, while the lower dots indicate the components of each group. Each bar corresponds to a compartment, with its height indicating the number of samples. As depicted in Fig. 2(b), it is evident that most samples are associated with more than one subcellular location. The number of miRNAs with subcellular localization in various compartments is as follows: 445 for nucleus, 232 for exosome, 229 for cytoplasm, and 79 for microvesicle, as shown

in Fig. 2(b). Each bar, from bottom to top, represents the number of miRNAs with four localization annotations down to those with a unique localization label. The total number of miRNAs in each compartment is labeled at the top of the bars in Fig. 2(b). Additionally, we analyzed the distribution of sequence lengths across four subcellular localizations, as illustrated in Fig. 2(c), with detailed sequence length distributions for all datasets across various subcellular locations provided in Fig. S1 and Fig. S2.

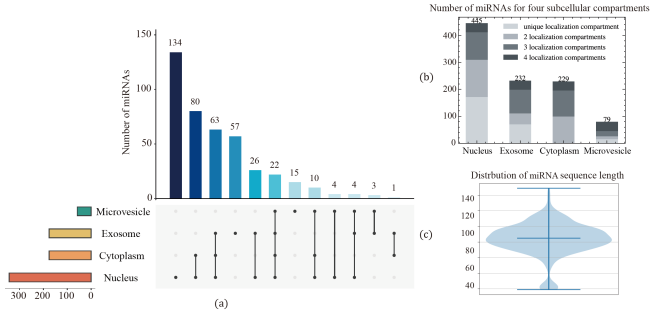

Fig 2. Distribution of miRNA subcellular localization dataset. (a) The numbers of miRNA subcellular localizations in the constructed dataset. (b) The statistics of miRNAs for four subcellular compartments. (c) The sequence length distribution of the collected dataset.

2) Training Datasets and Independent Test sets: The collected datasets were randomly split into training and test sets with an 8:2 ratio. The training set was used to train the subcellular localization classifier, while the test set was employed to evaluate its performance against the established predictor. Statistical summaries for both the benchmark dataset and the test set are presented in Tables I and II. Additionally, we quantified the sample sizes of instances exhibiting multiple concurrent labels within each location category.

TABLE I
STATISTICAL OF THE BENCHMARK DATASET

Subcellular	Conts	1	2	3	4	OAS
Nucleus	343	134	110	77	22	420
Exosome	177	58	30	67	22	
Cytoplasm	176	0	81	73	22	
Microvesicle	58	15	7	14	22	

Notes: Overall Actual Samples(OAS)

TABLE II
STATISTICAL OF THE INDEPENDENT TEST SET

Subcellular	Conts	1	2	3	4	OAS
Nucleus	102	38	28	25	11	118
Exosome	55	13	10	21	11	
Cytoplasm	53	0	19	23	11	
Microvesicle	21	1	3	6	11	

Notes: Overall Actual Samples(OAS)

C. MMLmiRLocNet architecture

In this work, Fig. 3 illustrates the architecture of MMLmiRLocNet, comprising four main components: (i) RNA embedding layer; (ii) Multi-view feature representations extract layer; (iii) a fully connected layer, and (iv) the multi-label classification layer.

MMLmiRLocNet takes a miRNA sequence as input, encoding lexical, syntactic, and semantic features for biological sequences. Following the feature embedding blocks, the network employs a multi-view feature extraction module to learn distinct weights for nucleotides associated with each subcellular localization. Lastly, a fully connected layer is employed to perform the multi-label classification task.

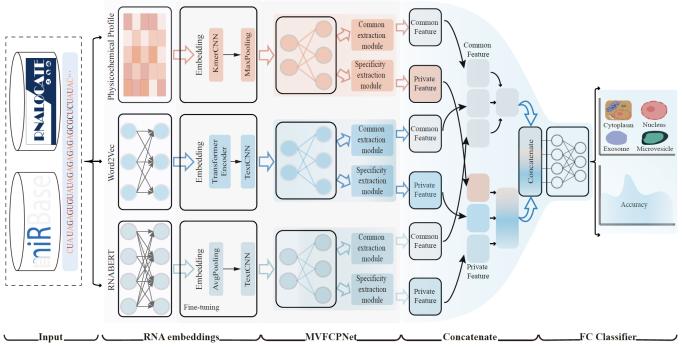

Fig 3. The framework of MMLmiRLocNet.

D. miRNA sequence input and embedding

In biological sequence processing, akin to natural language, sequences exhibit a range of complementary features [63]. To capture these diverse aspects, we collected a comprehensive array of feature representations that include lexical, syntactic, and semantic information. Before inputting raw miRNA into the prediction model, it must be converted into feature representations. In this study, we integrated lexical attributes derived from k-mer physicochemical profiles, syntactic characteristics obtained through word2vec embeddings, and semantic representations generated by the pre-trained BERT model RNABERT [64].

Given a benchmark dataset S for the prediction of miRNA subcellular localization task with N sequences:

$$S = \{B_1, B_2, B_3, B_4, \dots, B_i, \dots, B_N\} \quad (1)$$

where B_i is the i-th subsequence and S is the number of datasets.

The k-mer descriptor provides a straightforward approach for representing RNA sequences by quantifying the frequencies of k-neighboring nucleotides [65]. This method has proven effective in predicting human gene regulatory sequences and identifying enhancers [66, 67]. Furthermore, it facilitates the description of lexical representations of RNA sequences. For instance, the k-mer (k = 3) descriptor is calculated as follows:

$$V_{lexical} = \frac{N_t}{N} t \in (AAA, AAC, \dots\dots, UGC) \quad (2)$$

where N_t represents the count of k-mer type t, while N denotes the length of the sequence.

The word2vec descriptor computes a distributed word vector for each word within its given corpus context [68]. This word vector representation effectively portrays the semantics of

words. Within the word2vec model framework, consecutive k-mer nucleotide sequences in a window are treated as individual words and converted into feature vectors. The model was trained to encode a sequence window of 120 nucleotides into a 256-dimensional matrix. The maximum sequence length was fixed at 120 residues. In particular, sequences shorter than 120 nucleotides were zero-padded. The skip-gram approach of word2vec model is described as follows:

$$h = Wx_k \tag{3}$$

$$V_{syntactic} = W'h \tag{4}$$

where x_k represents the input for the target word; $V_{syntactic}$ denotes the output for the context words, and h denotes the hidden representation with W and W' denote distinct weights.

The RNABERT is adopted to effectively embed RNA bases to acquire semantically rich feature representations [69]. The maximum feature representation length was set to 105, aligning with the average length of miRNA sequences in the training dataset. RNA base embeddings can adequately capture the structural differences among RNA families. The RNABERT feature embedding is detailed as follows:

$$V_{semantic} = RNABERT(S) \tag{5}$$

where $RNABERT(*)$ is the pre-trained algorithm model; $V_{semantic}$ is the output representing the context words.

E. Multi-view feature extract network

In our approach, we utilize multi-view feature extraction methods that integrate various perspectives of data to derive comprehensive features. These methods incorporate nucleotide distribution representation of k-mer derived from RNA sequences, syntactic vector representation that captures the semantics of each word, and structural embedding representation. Each feature within the view serves a distinct role for prediction. For a given sequence, suppose the multi-view X with N samples set can be described as follows:

$$\{V^x \in \mathbb{R}^{n*d}\}_{x=1}^{X} \tag{6}$$

where V^x denotes the $x-$th view feature space, which are the inputs of MMLmiRLocNet.

Here, we propose a feature interaction module designed to extract consensus features and private features from heterogeneous data, respectively. This process involves implementing feature grafting during the feature learning phase. The multi-view consensus features and specific features extract network (MVFCPNet) is represented as follows:

$$\begin{cases} V_x^{\text{private}} = \text{ReLU}(\boldsymbol{W}_{private}V^x + \boldsymbol{b}_{private}) \\ V_x^{\text{comm}} = \text{ReLU}(\boldsymbol{W}_{comm}V^x + \boldsymbol{b}_{comm}) \end{cases} \tag{7}$$

where $\boldsymbol{W}_{private}$ and $\boldsymbol{b}_{private}$, $\boldsymbol{W}_{comm}$ and $\boldsymbol{b}_{comm}$ denotes private features or common feature weight matrices and bias vectors, respectively. V_{private}^x and V_{comm}^x are the x-th view private features and common features passed through the private feature extraction layer and the common feature extraction layer, respectively.

E. Fully connected and classification

In MMLmiRLocNet, distinct weights for private features contributing differently to the prediction of subcellular locations, and final feature vector is formed by concatenating both private and common features. These final features are then processed through a fully connected predictor layer:

$$\boldsymbol{F}_{final} = concat\left[V_{lexical}^{private}, V_{syntactic}^{private}, V_{semantic}^{private}, \boldsymbol{C}_{comm}\right] \tag{8}$$

$$\boldsymbol{y}_{pred} = \sigma\left(\boldsymbol{W}_{pred}\boldsymbol{F}_{final} + \boldsymbol{b}_{pred}\right) \tag{9}$$

where σ denotes the sigmoid activation function, $\mathbf{y}_{\text{pred}}$ are the predicted labels. In addition, the add network is utilized to extract supplementary information from common features across diverse representations of biological sequences, represented as:

$$\boldsymbol{C}_{comm} = \frac{1}{3}\sum_{x} V_x^{comm} \tag{10}$$

F. Performance evaluation

To assess model performance, five multi-label learning evaluation metrics are employed [70] including F1, $\text{AUC}_{\text{macro}}$ (AucMacro), Hamming Loss (Hloss), Subset Accuracy (SubsetAcc), and Accuracy. In cases of label imbalance, the F1 is critical as it integrates precision and recall into a single metric, offering a balanced assessment of a model's ability to identify relevant instances. Conversely, Hloss is essential for tasks requiring precise control over label errors, it measures the proportion of incorrect labels relative to the total number of labels, providing insight into error rates in multi-label classification. Additionally, subAcc is vital for scenarios requiring exact label matching. This metric assesses the proportion of samples where the predicted label set matches the true labels exactly, ensuring complete alignment between model predictions and actual labels.

Let $\text{D} = \{(x_i, \mathbb{Y}_i)|1 \le i \le p\}$ denotes p dimensional miRNA test space. Here, the x_i and $\mathbb{Y}_i = \{\mathbb{y}_{i,1}, \mathbb{y}_{i,2}, \cdots, \mathbb{y}_{i,q}\}$ represent the sequence feature set and possible label set of the miRNA sequence x_i, respectively. The metrics used for evaluation are defined as follows:

$$Accuracy(f) = \frac{1}{p}\sum_{i=1}^{p}\frac{\left|Y_i \cap h_{x_i}\right|}{\left|Y_i \cup h_{x_i}\right|} \tag{11}$$

$$\downarrow Hloss(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{q}\left|h(x_i)\Delta Y_i\right| \tag{12}$$

$$SubsetAcc(h) = \frac{1}{p}\sum_{i=1}^{M}\left[h\left(x^{(i)}\right) = y^{(i)}\right] \tag{13}$$

$$F1(h) = \frac{2 \cdot Precision(h) \cdot Recall(h)}{Precision(h) + Recall(h)} \tag{14}$$

$$\text{AucMacro}(h) = \frac{1}{p}\sum_{i=1}^{p}\text{AUC}_i \tag{15}$$

where AUC_i denotes the score for label i,Y_i represents the RNA subcellular location label set associated with $x_i, and\ \bar{Y}_i$ denotes its complement in $\mathbb{Y}_i$, indicating the label is not associated with x_i. The function $h(\cdot,\cdot)$ denotes label classifier. The lower Hloss rates indicate better model performance, and the greater subAcc, Accuracy, F1, and AucMacro values suggest the greater predictive capability of the model.

III. Results and Discussion

A. Comparative Performance with state-of-art predictors on test sets

In this section, we employed 10-fold cross-validation to determine optimal parameters of MMLmiRLocNet. We assessed its performance by comparing it with some proposed the state-of-the-art predictors of miRNA subcellular localizations using an independent test set. Consequently, we selected comparison methods based on specific criteria: (a) availability of an online web server or stand-alone package, (b) evaluation using a multi-label evaluation matrix, and (c) ability to predict miRNA subcellular localization using only sequences. As a consequence, MirLocPredictor, TextRNN, ncRNALocate-EL, MKSVM, and MK-GHkNN were selected. The detailed performances of these methods are presented in Table III. The hyperparameter settings are listed in Supplementary Table S4.

According to Table III, MMLmiRLocNet exhibits superior performance compared to the other predictors in terms of F1, subACC, and Accuracy, while also achieving the lowest Hloss. Therefore, our proposed MMLmiRLocNet outperforms all the other competing methods. These results demonstrate its strong capability in predicting multi-label miRNA subcellular localizations, showcasing its excellent performance on the test set.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON THE $\mathbb{Y}_{TEST}$

Methods	Acc	AucM	subAcc	F1	↓Hloss
MMLmiRLoc	**0.6124**	0.5953	**0.3152**	**0.7094**	**0.2627**
MirLocPredictor	0.5939	0.5758	0.2932	0.6933	0.2720
TextRNN	0.5504	0.5137	0.2186	0.6614	0.3146
ncRNALocate-EL	0.4548	**0.6091**	0.1271	0.5629	0.4047
MKSVM	0.5381	0.5562	0.2881	0.6313	0.2966
MK-GHkNN	0.5191	0.5598	0.2288	0.6172	0.3136

Notes: The numbers in bold indicate the best performance. Accuracy(Acc); AucMacro(AucM); SubsetAcc(subAcc); Hamming-loss(Hloss).

B. Performance at different subcellular locations

The predictive performance of the model across various subcellular locations was assessed by validating its performance on the test dataset. Evaluation metrics included the average area under the receiver operating characteristic curve (ROC-AUC) and the average area under the precision-recall curve (PR-AUC). The ROC-AUC and PR-AUC for MMLmiRLocNet across different subcellular locations are depicted separately in Fig. 4. Furthermore, the ROC curve and PR curve on the testing set are shown in Supplementary Fig. S3.

The results shown in Fig. 4 indicate that the ROC-AUC values for nucleus, cytoplasm, exosome, and microvesicle are 0.814, 0.692, 0.580, and 0.448, respectively. The PR-AUC of nucleus, cytoplasm, exosome, and microvesicle are 0.937, 0.597, 0.556, and 0.147, respectively. The results indicate that the proposed model exhibits better performance across all datasets, except for the microvesicle category on the testing set, which suggests that the model performs well in predicting subcellular locations for most samples.

C. Ablation study

To assess the impact of different feature components of MMLmiRLocNet on its prediction performance, we conducted an ablation study. We tested the model without MVFCPNet framework and assessed the performance of MVFCPNet framework with only two views. We retrained our model by sequentially adding individual feature components: (a) Lexical feature representations using k-mer descriptor, (b) Syntactic feature representations using word2vec embeddings, (c) Semantic feature representations using pre-trained algorithm RNABERT, and (d) the MVFCPNet framework with common feature and the private feature. Detailed experimental results are provided in Table IV.

According to the results, we observe that the MVFCPNet framework constitutes a crucial component of MMLmiRLocNet. Without MVFCPNet, Accuracy, AucMacro, subACC, F1, and Hloss decrease to 0.5836, 05725, 0.2771, 0.6870, and 0.2835, respectively. Furthermore, as feature modules were incrementally added, the model demonstrated progressively improved performance in predicting subcellular locations. The proposed approach performs effectively by leveraging lexical, syntactic, and semantic embeddings from RNA feature representations and multi-view feature extraction methods, representing a notable advantage in our prediction method. In conclusion, this achievement can be attributed not only to the model architecture but also to the other critical factors. In addition, we tested the importance of common features in MMLmiRLocNet, as its shown in Supplementary Fig. S4, which shows the contribution of common features to the overall performance.

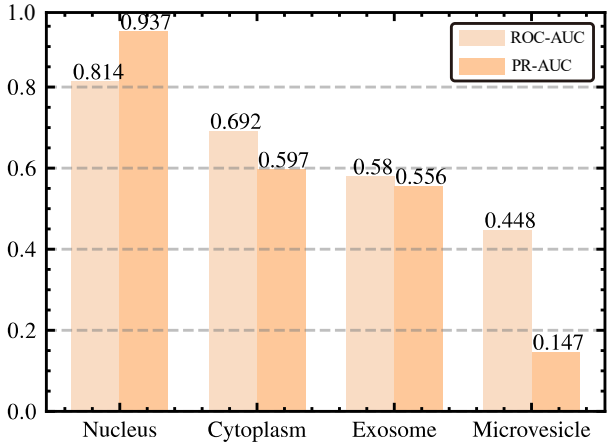

Fig 4. ROC-AUC and PR-AUC score on $\mathbb{Y}_{test}$

D. The Effectiveness of Bert Size

To better understand the role of the pre-trained algorithm, the impact of RNABERT embedding feature dimension sizes on model performance is illustrated in Table V. We conducted tests using randomly initialized embedded feature dimensions of sizes 32, 64, 105, 128, 256, and 440, while keeping the remainder of the model constant. This approach allows us to

examine how varying the dimensionality affects the model's ability to capture and generalize RNA sequence features.

As is shown in Table V, the classification models achieved optimal performance across all evaluation measures with a dimension size of 105. However, performance declines with dimension sizes exceeding 105. The reason is that a dimension size that is too small may truncate crucial sequence information needed for accurate prediction. Conversely, excessively large output dimensions may lead RNABERT to incorporate redundant information, complicating the prediction of subcellular locations.

TABLE IV
THE QUANTITATIVE RESULTS OF ABLATION STUDIES ON DIFFERENT MODEL COMPONENTS

Feature Ablation	Acc	AucM	subAcc	F1	↓Hloss
Kmer	0.5433	0.5421	0.2474	0.6443	0.3146
Word2Vec	0.5741	0.5657	0.2814	0.6802	0.2828
RNAbert	0.5787	0.5668	0.2771	0.6856	0.2828
Without MVL	0.5836	05725	0.2771	0.6870	0.2835
Only RNAbert and Word2Vec with MVL	0.5864	0.5733	0.3042	0.6856	0.2779
MMLmiRLocNet	**0.6124**	**0.5953**	**0.3152**	**0.7094**	**0.2627**

Notes: The numbers in bold indicate the best performance; Kmer, use k-mer 4 feature without multi-view multi-label learning approach; Word2Vec, use word2vec method to encode RNA sequence information without multi-view multi-label learning approach; RNAbert, use pre-trained RNABERT model to encode RNA sequence information without multi-view multi-label learning approach; MVL, Multi-view Learning Approach. Accuracy(Acc); AucMacro(AucM); SubsetAcc(subAcc); Hamming-loss(Hloss).

TABLE V
PERFORMANCE COMPARISON OF RNABERT HYPERPARAMETERS ON THE PREDICTIVE MODEL ON PERFORMANCE

Bert Size	Acc	AucM	subAcc	F1	↓Hloss
32	0.5815	0.5732	0.2797	0.6818	0.2871
64	0.598	0.5891	0.3110	0.6966	0.2699
105	**0.6124**	**0.5953**	**0.3152**	**0.7094**	**0.2627**
128	0.6044	0.5881	0.3093	0.7028	0.2678
256	0.6064	0.5864	0.3068	0.7002	0.2638
440	0.5949	0.5817	0.3051	0.6924	0.2718

Notes: The numbers in bold indicate the best performance. Accuracy(Acc); AucMacro(AucM); SubsetAcc(subAcc); Hamming-loss(Hloss).

E. Web server

To enhance accessibility to MMLmiRLocNet, a user-friendly web server has been developed, and is available at http://bliulab.net/MMLmiRLocNet. Researchers can input the miRNA sequence into the provided box and submit it to view the prediction results. The server backend automatically performs feature extraction on the input sequences and loads the well-trained model for prediction. MMLmiRLocNet is highly regarded for its convenience and efficiency in predicting miRNA subcellular localization.

VI. CONCLUSION

In this study, we introduce a novel computational approach, named MMLmiRLocNet, designed for the accurate prediction of miRNA subcellular localization using a multi-view multi-label learning framework, which is crucial for drug design. This method leverages multiple perspectives of miRNA data to enhance prediction robustness and accuracy across various subcellular compartments. Compared to previous research, MMLmiRLocNet has two primary innovations: (i) extracting diverse feature representations of biological sequences by investigating lexical, syntactic, and semantic aspects, and (ii) employing consensus and specific feature extraction networks inspired by the multi-view multi-label learning method to identify multiple subcellular localizations. Experimental results show that MMLmiRLocNet outperforms the state-of-the-art methods. The ablation study analysis validated that the consensus and specific feature extraction networks contribute to the prediction precision of miRNA subcellular localization. The proposed method provides a fresh perspective on the subcellular localization of various RNA types.

In future research, the following directions could be explored:

(i) The current absence of adequate baseline datasets poses challenges for researchers in systematically constructing accurate models for predicting miRNA subcellular locations. Integrating miRNA sequence information with other relevant multi-source biological datasets [71], such as interaction networks involving miRNAs and lncRNAs [72], would be beneficial.

(ii) In this study, we initially explored using the language model RNABERT to construct feature representations without using the effective characterization of structural information. As large-scale language models and pre-trained models continue to advance rapidly [73, 74], integrating structural information into these models for RNA function analysis [75], holds promise for enhancing their effectiveness in RNA function prediction tasks.

REFERENCES

[1] P. J. Dexheimer, and L. Cochella, "MicroRNAs: from mechanism to organism," *Frontiers in cell and developmental biology,* vol. 8, pp. 409, 2020.

[2] X. Chen, and O. Rechavi, "Plant and animal small RNA communications between cells and organisms," *Nature Reviews Molecular Cell Biology,* vol. 23, no. 3, pp. 185-203, 2022.

[3] L. Jiao, Y. Liu, X.-Y. Yu, X. Pan, Y. Zhang, J. Tu, Y.-H. Song, and Y. Li, "Ribosome biogenesis in disease: new players and therapeutic targets," *Signal*

Transduction and Targeted Therapy, vol. 8, no. 1, pp. 15, 2023.

[4] C. Ding, H. Xu, Z. Yu, M. Roulis, R. Qu, J. Zhou, J. Oh, J. Crawford, Y. Gao, and R. Jackson, "RNA m6A demethylase ALKBH5 regulates the development of γδ T cells," *Proceedings of the National Academy of Sciences,* vol. 119, no. 33, pp. e2203318119, 2022.

[5] R. Shang, S. Lee, G. Senavirathne, and E. C. Lai, "microRNAs in action: biogenesis, function and regulation," *Nature Reviews Genetics,* vol. 24, no. 12, pp. 816-833, 2023.

[6] L. F. Gebert, and I. J. MacRae, "Regulation of microRNA function in animals," *Nature reviews Molecular cell biology,* vol. 20, no. 1, pp. 21-37, 2019.

[7] M. Jie, T. Feng, W. Huang, M. Zhang, Y. Feng, H. Jiang, and Z. Wen, "Subcellular localization of miRNAs and implications in cellular homeostasis," *Genes,* vol. 12, no. 6, pp. 856, 2021.

[8] J. Li, Q. Zou, and L. Yuan, "A review from biological mapping to computation-based subcellular localization," *Molecular Therapy-Nucleic Acids,* vol. 32, pp. 507-521, 2023.

[9] A. F. Savulescu, E. Bouilhol, N. Beaume, and M. Nikolski, "Prediction of RNA subcellular localization: learning from heterogeneous data sources," *Iscience,* vol. 24, no. 11, 2021.

[10] Y. Chen, and X. Wang, "miRDB: an online database for prediction of functional microRNA targets," *Nucleic acids research,* vol. 48, no. D1, pp. D127-D131, 2020.

[11] J. Rivera, L. Gangwani, and S. Kumar, "Mitochondria localized microRNAs: an unexplored miRNA niche in Alzheimer's disease and aging," *Cells,* vol. 12, no. 5, pp. 742, 2023.

[12] R. Li, H. Qu, S. Wang, J. M. Chater, X. Wang, Y. Cui, L. Yu, R. Zhou, Q. Jia, and R. Traband, "CancerMIRNome: an interactive analysis and visualization database for miRNome profiles of human cancer," *Nucleic acids research,* vol. 50, no. D1, pp. D1139-D1146, 2022.

[13] T. J. Lee, X. Yuan, K. Kerr, J. Y. Yoo, D. H. Kim, B. Kaur, and H. K. Eltzschig, "Strategies to modulate microRNA functions for the treatment of cancer or organ injury," *Pharmacological Reviews,* vol. 72, no. 3, pp. 639-667, 2020.

[14] W. Tang, S. Wan, Z. Yang, A. E. Teschendorff, and Q. Zou, "Tumor origin detection with tissue-specific miRNA and DNA methylation markers," *Bioinformatics,* vol. 34, no. 3, pp. 398-406, Feb 1, 2018.

[15] M. Liu, C. Li, R. Chen, D. Cao, and X. J. E. S. w. A. Zeng, "Geometric Deep Learning for Drug Discovery," *Expert Systems with Applications*, pp. 122498, 2023.

[16] C. E. Condrat, D. C. Thompson, M. G. Barbu, O. L. Bugnar, A. Boboc, D. Cretoiu, N. Suciu, S. M. Cretoiu, and S. C. Voinea, "miRNAs as biomarkers in disease: latest findings regarding their role in diagnosis and prognosis," *Cells,* vol. 9, no. 2, pp. 276, 2020.

[17] H. Walgrave, L. Zhou, B. De Strooper, and E. Salta, "The promise of microRNA-based therapies in Alzheimer's disease: challenges and perspectives," *Molecular neurodegeneration,* vol. 16, pp. 1-16, 2021.

[18] L. Valihrach, P. Androvic, and M. Kubista, "Circulating miRNA analysis for cancer diagnostics and therapy," *Molecular Aspects of Medicine,* vol. 72, pp. 100825, 2020.

[19] G. P. Brennan, and D. C. Henshall, "MicroRNAs as regulators of brain function and targets for treatment of epilepsy," *Nature Reviews Neurology,* vol. 16, no. 9, pp. 506-519, 2020.

[20] Q. Zou, P. Xing, L. Wei, and B. Liu, "Gene2vec: gene subsequence embedding for prediction of mammalian N-6-methyladenosine sites from mRNA," *Rna,* vol. 25, no. 2, pp. 205-218, Feb, 2019.

[21] T. Bai, and B. Liu, "ncRNALocate-EL: a multi-label ncRNA subcellular locality prediction model based on ensemble learning," *Briefings in Functional Genomics,* vol. 22, no. 5, pp. 442-452, 2023.

[22] H. Zhou, H. Wang, J. Tang, Y. Ding, and F. Guo, "Identify ncRNA subcellular localization via graph regularized k-local hyperplane distance nearest neighbor model on multi-kernel learning," *IEEE/ACM Trans Comput Biol Bioinform,* vol. PP, Aug 25, 2021.

[23] H. Wang, Y. Ding, J. Tang, Q. Zou, and F. Guo, "Identify RNA-associated subcellular localizations based on multi-label learning using Chou's 5-steps rule," *BMC Genomics,* vol. 22, no. 1, pp. 56, Jan 15, 2021.

[24] H. Cui, J. Zhai, and C. Ma, "miRLocator: machine learning-based prediction of mature microRNAs within plant pre-miRNA sequences," *PLoS One,* vol. 10, no. 11, pp. e0142753, 2015.

[25] M. N. Asim, M. I. Malik, C. Zehe, J. Trygg, A. Dengel, and S. Ahmed, "MirLocPredictor: a ConvNet-based multi-label MicroRNA subcellular localization predictor by incorporating k-Mer positional information," *Genes,* vol. 11, no. 12, pp. 1475, 2020.

[26] M. N. Asim, M. A. Ibrahim, C. Zehe, O. Cloarec, R. Sjogren, J. Trygg, A. Dengel, and S. Ahmed, "L2S-MirLoc: a lightweight two stage MiRNA sub-cellular localization prediction framework." pp. 1-8.

[27] P. K. Meher, S. Satpathy, and A. R. Rao, "miRNALoc: predicting miRNA subcellular localizations based on principal component scores of physico-chemical properties and pseudo compositions of di-nucleotides," *Scientific Reports,* vol. 10, no. 1, pp. 14557, 2020.

[28] Z. Liu, T. Bai, B. Liu, and L. Yu, "MulStack: An ensemble learning prediction model of multilabel mRNA subcellular localization," *Computers in Biology and Medicine,* vol. 175, pp. 108289, 2024.

[29] Z.-Y. Zhang, L. Ning, X. Ye, Y.-H. Yang, Y. Futamura, T. Sakurai, and H. Lin, "iLoc-miRNA: extracellular/intracellular miRNA prediction using deep BiLSTM with attention mechanism," *Briefings in Bioinformatics,* vol. 23, no. 5, pp. bbac395, 2022.

[30] M. Zeng, Y. Wu, C. Lu, F. Zhang, F.-X. Wu, and M. Li, "DeepLncLoc: a deep learning framework for long

non-coding RNA subcellular localization prediction based on subsequence embedding," *Briefings in Bioinformatics,* vol. 23, no. 1, pp. bbab360, 2022.

[31] Q. Zou, P. Xing, L. Wei, and B. Liu, "Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA," *Rna,* vol. 25, no. 2, pp. 205-218, 2019.

[32] Q. Zou, P. Xing, L. Wei, and B. Liu, "Gene2vec: Gene Subsequence Embedding for Prediction of Mammalian N6-Methyladenosine Sites from mRNA," *RNA,* vol. 25, no. 2, pp. 205-218, 2019.

[33] Z. Chen, P. Zhao, C. Li, F. Li, D. Xiang, Y.-Z. Chen, T. Akutsu, R. J. Daly, G. I. Webb, and Q. Zhao, "iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization," *Nucleic acids research,* vol. 49, no. 10, pp. e60-e60, 2021.

[34] B. Liu, X. Gao, and H. Zhang, "BioSeq-Analysis2. 0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches," *Nucleic acids research,* vol. 47, no. 20, pp. e127-e127, 2019.

[35] Y. Yang, X. Fu, W. Qu, Y. Xiao, and H.-B. Shen, "MiRGOFS: a GO-based functional similarity measurement for miRNAs, with applications to the prediction of miRNA subcellular localization and miRNA–disease association," *Bioinformatics,* vol. 34, no. 20, pp. 3547-3556, 2018.

[36] M. Xu, Y. Chen, Z. Xu, L. Zhang, H. Jiang, and C. Pian, "MiRLoc: predicting miRNA subcellular localization by incorporating miRNA–mRNA interactions and mRNA subcellular localization," *Briefings in Bioinformatics,* vol. 23, no. 2, pp. bbac044, 2022.

[37] T. Bai, K. Yan, and B. Liu, "DAmiRLocGNet: miRNA subcellular localization prediction by combining miRNA–disease associations and graph convolutional networks," *Briefings in Bioinformatics,* vol. 24, no. 4, pp. bbad212, 2023.

[38] Y. Liu, X. Shen, Y. Gong, Y. Liu, B. Song, and X. J. B. i. B. Zeng, "Sequence Alignment/Map format: a comprehensive review of approaches and applications," *Briefings in Bioinformatics,* vol. 24, no. 5, pp. bbad320, 2023.

[39] J. Qiao, J. Jin, H. Yu, and L. Wei, "Towards Retraining-free RNA Modification Prediction with Incremental Learning," *Information Sciences*, pp. 120105, 2024.

[40] F. Wang, and L. Wei, "Multi-scale deep learning for the imbalanced multi-label protein subcellular localization prediction based on immunohistochemistry images," *Bioinformatics,* vol. 38, no. 9, pp. 2602-2611, 2022.

[41] D. Wang, Z. Zhang, Y. Jiang, Z. Mao, D. Wang, H. Lin, and D. Xu, "DM3Loc: multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism," *Nucleic acids research,* vol. 49, no. 8, pp. e46-e46, 2021.

[42] X. Guo, W. Zhou, B. Shi, X. Wang, A. Du, Y. Ding, J. Tang, and F. Guo, "An Efficient Multiple Kernel Support Vector Regression Model for Assessing Dry Weight of Hemodialysis Patients," *Current Bioinformatics,* vol. 16, no. 2, pp. 284-293, 2021, 2021.

[43] M. Zeng, Y. Wu, Y. Li, R. Yin, C. Lu, J. Duan, and M. Li, "LncLocFormer: a Transformer-based deep learning model for multi-label lncRNA subcellular localization prediction by using localization-specific attention mechanism," *Bioinformatics,* vol. 39, no. 12, pp. btad752, 2023.

[44] H. Zhou, H. Wang, J. Tang, Y. Ding, and F. Guo, "Identify ncRNA Subcellular Localization via Graph Regularized k-Local Hyperplane Distance Nearest Neighbor Model on Multi-Kernel Learning," *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 19, no. 6, pp. 3517-3529, 2022, 2022.

[45] Y. Cheng, Q. Li, Y. Wang, and W. Zheng, "Multi-view multi-label learning with view feature attention allocation," *Neurocomputing,* vol. 501, pp. 857-874, 2022.

[46] Z. Y. Zhang, Z. Zhang, X. Ye, T. Sakurai, and H. Lin, "A BERT-based model for the prediction of lncRNA subcellular localization in Homo sapiens," *Int J Biol Macromol,* vol. 265, no. Pt 1, pp. 130659, Mar 10, 2024.

[47] R. Wang, Y. Jiang, J. Jin, C. Yin, H. Yu, F. Wang, J. Feng, R. Su, K. Nakai, and Q. Zou, "DeepBIO: an automated and interpretable deep-learning platform for high-throughput biological sequence prediction, functional annotation and visualization analysis," *Nucleic Acids Research,* vol. 51, no. 7, pp. 3017-3029, 2023.

[48] M. Liang, Y. Zhan, and R. W. Liu, "MVFFNet: Multi-view feature fusion network for imbalanced ship classification," *Pattern Recognition Letters,* vol. 151, pp. 26-32, 2021.

[49] X. Qiang, C. Zhou, X. Ye, P.-f. Du, R. Su, and L. Wei, "CPPred-FL: a sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning," *Briefings in Bioinformatics,* vol. 21, no. 1, pp. 11-23, 2020.

[50] N. D. Nguyen, and D. Wang, "Multiview learning for understanding functional multiomics," *PLoS computational biology,* vol. 16, no. 4, pp. e1007677, 2020.

[51] Y. Wang, Zhai, Y., Ding, Y., Zou, Q, "SBSM-Pro: Support Bio-sequence Machine for Proteins," *arXiv preprint*, pp. arXiv:2308.10275, 2023.

[52] H. Zulfiqar, Z. Guo, R. M. Ahmad, Z. Ahmed, P. Cai, X. Chen, Y. Zhang, H. Lin, and Z. Shi, "Deep-STP: a deep learning-based approach to predict snake toxin proteins by using word embeddings," *Frontiers in Medicine,* vol. 10, 2024-January-17, 2024.

[53] C. Liang, L. Wang, L. Liu, H. Zhang, and F. Guo, "Multi-view unsupervised feature selection with tensor robust principal component analysis and

consensus graph learning," *Pattern Recognition,* vol. 141, Sep, 2023.

[54] H. Han, B. A. Talpur, W. Liu, L. Wang, B. Ahmed, N. Sarhan, and E. M. Awwad, "RNA-RBP interactions recognition using multi-label learning and feature attention allocation," *Journal of Cloud Computing,* vol. 13, no. 1, pp. 54, 2024.

[55] C. Ao, X. Ye, T. Sakurai, Q. Zou, and L. Yu, "m5U-SVM: identification of RNA 5-methyluridine modification sites based on multi-view features of physicochemical features and distributed representation," *BMC biology,* vol. 21, no. 1, pp. 93, 2023.

[56] F. Sun, H. Xu, Y. Meng, Z. Lu, and C. Gong, "BERT-based coupling evaluation of biological strategies in bio-inspired design," *Expert Systems with Applications,* vol. 222, pp. 119725, 2023.

[57] T. Bepler, and B. Berger, "Learning the protein language: Evolution, structure, and function," *Cell systems,* vol. 12, no. 6, pp. 654-669. e3, 2021.

[58] H. Y. Zhang, Q. Zou, Y. Ju, C. G. Song, and D. Chen, "Distance-based Support Vector Machine to Predict DNA N6-methyladenine Modification," *Current Bioinformatics,* vol. 17, no. 5, pp. 473-482, 2022.

[59] W. Zhu, S. S. Yuan, J. Li, C. B. Huang, H. Lin, and B. Liao, "A First Computational Frame for Recognizing Heparin-Binding Protein," *Diagnostics (Basel),* vol. 13, no. 14, Jul 24, 2023.

[60] T. Cui, Y. Dou, P. Tan, Z. Ni, T. Liu, D. Wang, Y. Huang, K. Cai, X. Zhao, and D. Xu, "RNALocate v2. 0: an updated resource for RNA subcellular localization with increased coverage and annotation," *Nucleic acids research,* vol. 50, no. D1, pp. D333-D339, 2022.

[61] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, "miRBase: from microRNA sequences to function," *Nucleic acids research,* vol. 47, no. D1, pp. D155-D162, 2019.

[62] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT Suite: a web server for clustering and comparing biological sequences," *Bioinformatics,* vol. 26, no. 5, pp. 680-682, 2010.

[63] H.-L. Li, Y.-H. Pang, and B. Liu, "BioSeq-BLM: a platform for analyzing DNA, RNA and protein sequences based on biological language models," *Nucleic acids research,* vol. 49, no. 22, pp. e129-e129, 2021.

[64] M. Akiyama, and Y. Sakakibara, "Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning," *NAR genomics and bioinformatics,* vol. 4, no. 1, pp. lqac012, 2022.

[65] C. Lorenzi, S. Barriere, J.-P. Villemin, L. Dejardin Bretones, A. Mancheron, and W. Ritchie, "iMOKA: k-mer based software to analyze large collections of sequencing data," *Genome biology,* vol. 21, pp. 1-19, 2020.

[66] L. Cai, X. Ren, X. Fu, L. Peng, M. Gao, and X. Zeng, "iEnhancer-XG: interpretable sequence-based enhancers and their strength predictor," *Bioinformatics,* vol. 37, no. 8, pp. 1060-1067, 2021.

[67] A. Field, and K. Adelman, "Evaluating enhancer function and transcription," *Annual review of biochemistry,* vol. 89, no. 1, pp. 213-234, 2020.

[68] H. Kurata, S. Tsukiyama, and B. Manavalan, "iACVP: markedly enhanced identification of anti-coronavirus peptides using a dataset-specific word2vec model," *Briefings in bioinformatics,* vol. 23, no. 4, pp. bbac265, 2022.

[69] K. Hara, N. Iwano, T. Fukunaga, and M. Hamada, "DeepRaccess: high-speed RNA accessibility prediction using deep learning," *Frontiers in Bioinformatics,* vol. 3, pp. 1275787, 2023.

[70] M.-L. Zhang, and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE transactions on knowledge and data engineering,* vol. 26, no. 8, pp. 1819-1837, 2013.

[71] L.-X. Guo, Z.-H. You, L. Wang, C.-Q. Yu, B.-W. Zhao, Z.-H. Ren, and J. Pan, "A novel circRNA-miRNA association prediction model based on structural deep neural network embedding," *Briefings in Bioinformatics,* vol. 23, no. 5, pp. bbac391, 2022.

[72] J. Venkatesh, M.-C. D. Wasson, J. M. Brown, W. Fernando, and P. Marcato, "LncRNA-miRNA axes in breast cancer: Novel points of interaction for strategic attack," *Cancer letters,* vol. 509, pp. 81-88, 2021.

[73] B. Song, Z. Li, X. Lin, J. Wang, T. Wang, and X. Fu, "Pretraining model for biological sequence data," *Briefings in functional genomics,* vol. 20, no. 3, pp. 181-195, 2021.

[74] N. Wang, J. Bian, Y. Li, X. Li, S. Mumtaz, L. Kong, and H. Xiong, "Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning," *Nature Machine Intelligence*, pp. 1-10, 2024.

[75] C. V. Theodoris, L. Xiao, A. Chopra, M. D. Chaffin, Z. R. Al Sayed, M. C. Hill, H. Mantineo, E. M. Brydon, Z. Zeng, and X. S. Liu, "Transfer learning enables predictions in network biology," *Nature,* vol. 618, no. 7965, pp. 616-624, 2023.