M6ATMR: identifying N6-methyladenosine sites through RNA sequence similarity matrix reconstruction guided by Transformer

Shuang Xiang, Te Zhang and Minghao Wu

Changjiang Water Resources and Hydropower Development Group, Wuhan, Hubei, China

ABSTRACT

Numerous studies have focused on the classification of N6-methyladenosine (m6A) modification sites in RNA sequences, treating it as a multi-feature extraction task. In these studies, the incorporation of physicochemical properties of nucleotides has been applied to enhance recognition efficacy. However, the introduction of excessive supplementary information may introduce noise to the RNA sequence features, and the utilization of sequence similarity information remains underexplored. In this research, we present a novel method for RNA m6A modification site recognition called M6ATMR. Our approach relies solely on sequence information, leveraging Transformer to guide the reconstruction of the sequence similarity matrix, thereby enhancing feature representation. Initially, M6ATMR encodes RNA sequences using 3-mers to generate the sequence similarity matrix. Meanwhile, Transformer is applied to extract sequence structure graphs for each RNA sequence. Subsequently, to capture low-dimensional representations of similarity matrices and structure graphs, we introduce a graph self-correlation convolution block. These representations are then fused and reconstructed through the local-global fusion block. Notably, we adopt iteratively updated sequence structure graphs to continuously optimize the similarity matrix, thereby constraining the end-to-end feature extraction process. Finally, we employ the random forest (RF) algorithm for identifying m6A modification sites based on the reconstructed features. Experimental results demonstrate that M6ATMR achieves promising performance by solely utilizing RNA sequences for m6A modification site identification. Our proposed method can be considered an effective complement to existing RNA m6A modification site recognition approaches.

Subjects Bioinformatics, Data Mining and Machine Learning

Keywords RNA modification, N6-methyladenosine, Transformer, Similarity matrix, Graph

INTRODUCTION

To date, approximately 160 chemical modifications have been discerned in RNA, substantially enriching the diversity of RNA function and genetic information (*Rehman et al., 2021*). Among these modifications, N6-methyladenosine (m6A) stands out as the most prevalent modification type in eukaryotes and the sole dynamic, reversible RNA modification, along with N1-methyladenosine (*Wang & Yan, 2018*). m6A plays pivotal

Submitted 26 January 2023
Accepted 24 July 2023
Published 11 September 2023

Corresponding author
Minghao Wu, wu.mh@crhdc.com.cn

Academic editor
Liang Wang

Additional Information and Declarations can be found on page 15

DOI **10.7717/peerj.15899**

roles in various cellular processes, including cell growth, mRNA selective splicing, stem cell differentiation, and circadian clock control (*Fustin et al., 2013*; *Geula et al., 2015*; *Wang & Wang, 2020*; *Wang et al., 2014*; *Wang et al., 2018*). Furthermore, m6A exhibits close associations with the pathogenesis of diverse diseases, such as prostate cancer, acute myeloid leukemia, and thyroid tumors (*Rehman et al., 2021*). Given the significance of m6A, there exists an imperative to identify potential m6A modification sites.

High-throughput sequencing techniques have been extensively utilized for the identification of m6A modification sites, including m6A sequencing (*Dominissini et al., 2012*), crosslinking immunoprecipitation (*Ke et al., 2015*), and Methylated RNA Immunoprecipitation (*Meyer et al., 2012*). These methodologies have significantly contributed to our understanding of m6A modification on RNA. However, due to the dynamic and tissue-specific nature of m6A modification sites, limited experimental approaches may not be sufficiently flexible in identifying potential modification sites (*Wang & Yan, 2018*). Furthermore, the wet-lab experiments employed to identify m6A modification sites are often costly and time-consuming. In recent years, an increasing number of studies have recognized the notable advantages of computational methods for identifying RNA m6A modification sites. These computational approaches offer high generalization, rapid processing times, and lower costs, rendering them an attractive and viable alternative.

The computational identification of RNA m6A modification sites can be broadly categorized into two groups: machine learning-based methods and deep learning-based methods. Both approaches share a common core, which is the feature extraction task, aimed at capturing better representations of RNA sequences for improved recognition performance. Machine learning-based methods typically involve two main stages: feature engineering and downstream classification. In feature engineering, various coding approaches have been applied to represent RNA sequences, including k-mer, one-hot coding, accumulated nucleotide frequency (*Chen et al., 2015*), composition of k-space nucleic acid pairs (*Zhang et al., 2020*), dinucleotide composition (*Di Giallonardo et al., 2017*), and enhanced nucleic acid composition (*Huang et al., 2018*). The iRNA toolkits (*Chen et al., 2018*; *Qiu et al., 2017*; *Yang et al., 2018*) are noteworthy examples that utilize these encoding methods. Moreover, the iRNA toolkits were the pioneers in incorporating the physicochemical properties of nucleotides for recognizing various types of RNA modification sites. In the downstream classification task, different classifiers are usually employed to recognize the extracted features, including random forest (RF) (*Breiman, 2001*). The iRNA toolkits, AthMethPre (*Xiang et al., 2016*), M6ATH (*Chen et al., 2016*), and RAM-NPPS (*Xing et al., 2017*) prefer Support Vector Machine (SVM), while other methods like SRAMP (*Zhou et al., 2016*) and M6AMRFS (*Qiang et al., 2018*) explore ensemble methods with multiple classifiers for downstream tasks. On the other hand, deep learning-based methods view feature extraction and classification as continuous processes, allowing them to learn more semantic information from the data. Researchers are increasingly turning to deep learning strategies for RNA m6A modification site recognition tasks. For example, iN6-Methyl (*Nazari et al., 2019*) and m6AGE (*Wang et al., 2021*) use convolutional neural networks to extract sequence features.

These studies have made significant progress in identifying RNA m6A modification sites. However, they also have limitations. For instance, incorporating additional information, like the physicochemical properties of nucleotides, alongside RNA sequences may introduce potential information interference. Moreover, these methods have mainly focused on learning features from sequential nucleotide distributions, potentially overlooking associations of nucleotides through self-correlations in RNA sequences.
To address these issues, we propose a novel approach in this article, named m6ATMR, for RNA m6A modification site recognition. m6ATMR utilizes Transformer (*Vaswani et al., 2017*) to guide the reconstruction of the nucleotide similarity matrix, thereby enhancing feature representations of RNA sequences in a sequence-dependent manner. Specifically, RNA sequences are first encoded using the 3-mer method, generating the initial similarity matrix for each sequence. Then, Transformer is applied to further obtain the sequence structure graphs of RNA sequences. To optimize the sequence structure graphs, we calculate the Manhattan distance and perform threshold screening on the vector representation from Transformer. Next, we design a graph self-correlation convolution block to obtain low-dimensional representations of both the similarity matrix and the structure graph. In addition, we dynamically combine the low-dimensional representations obtained from the initial 3-mer representations of RNA sequences, considering both local and global perspectives, to create the final recombined features. To explore potential nucleotide associations in RNA sequences, we use iteratively updated sequence structure graphs to continuously optimize the similarity matrices, further enhancing the end-to-end feature extraction process. Finally, we employ the random forest (RF) algorithm to classify and recognize RNA sequences based on the learned features. By following this approach, m6ATMR aims to overcome the limitations of previous methods and improve the accuracy of RNA m6A modification site identification.

The main contributions of this article are as follows: First, we propose a sequence-based approach for identifying RNA m6A modification sites without introducing potentially misleading additional information. Second, the similarity matrices of RNA sequences are computed to provide more effective information that can be learned for sequences, and on this basis, the Transformer is used to reconstruct the similarity matrices and further optimize the sequence representations. Third, we propose a graph self-correlation convolution to learn a low-dimensional representation of the sequence without introducing prior information about the nodes. A series of experiments demonstrate the effectiveness of the representation strategies of M6ATMR. For M6ATMR, it is worth noting that relying solely on RNA sequences for m6A modification site identification can reduce the need for additional prior information, while still ensuring high identification accuracy. In addition, the sequence representations learned by our model perform consistently well across different classifiers, indicating that our method is not dependent on the choice of a specific classifier. In conclusion, our experimental results demonstrate that M6ATMR achieves excellent performance in identifying m6A modification sites using only RNA sequences. This highlights its effectiveness as a complementary method for RNA m6A modification site recognition.

MATERIALS AND METHODS

Problem description and datasets

One of the focuses of RNA modification research is site recognition. For this task, the computational methods are usually to convert the problem into a binary classification problem, which takes the RNA sequence information as the initial input to the classification model and gets the probability value of modification sites. The method in our article follows this paradigm. For a given sequence $X = \{x_1, x_2 \ldots x_n\}$, we summarize the model objective as $Y = f_\sigma(X) \in \{0, 1\}$, where $f_\sigma(\cdot)$ is the optimal mapping relationship, and Y is the prediction label. If the prediction result is a positive sample, $Y = 1$, otherwise $Y = 0$. It is worth noting that, as in most studies, the inclusion of m6A modification sites is used as the classification criteria for positive and negative samples. That is, for a given sequence X, if its central position is the modified site, sequence X is regarded as a positive sample. Negative samples do not contain modification sites. To this end, we select the RNA m6A modification site dataset of *Arabidopsis thaliana* (A101 dataset) (*Wan et al., 2015*) for study. The dataset contains 2,100 samples, in which the ratio of positive and negative samples is 1:1.

Model description

M6ATMR implements the identification task of m6A modification sites based on several procedures. First, the RNA sequences were processed into readable coding representation based on k-mer algorithm, and the similarity between RNA sequences was measured by Manhattan distance to further construct the similarity matrix. After that, M6ATMR learned the structural information of RNA sequences through the Transformer encoder, and converts this structural information into the structural graph by Manhattan distance. It then inputted the similarity matrix and structure graph into the self-correlation graph neural network to iteratively update the similarity matrix and obtain the low-dimensional representations of each RNA sequence. These low-dimensional representations were passed through a local-global fusion block to generate the final fusion representation, which was fed into the RF for identification. For the convenience of description, section 2 describes M6ATMR in detail from four parts: similarity matrix calculation, structure graph learning, similarity matrix optimization, and local-global representation fusion.
The framework of M6ATMR is shown in Fig. 1, and the details of each part are as follows.

Similarity matrix calculation

For similarity matrix calculation, the key step is to transform RNA sequences into the readable representations using k-mer frequency statistics. Statistical k-mer frequency information can reveal the distribution law of seed sequences and is an important tool to study sequence similarity. In sequence X, a substring of length K refers to K monometric units starting from any position in X, which is called k-mer. In this article, considering that each of the three adjacent nucleotides in an mRNA molecule is organized into a group that represents the pattern of a particular amino acid in protein synthesis, therefore, to make the model biological interpretation, we set K to 3. The algorithm requires that the starting position of the sequence set $\Omega = \{X_1, X_2, \ldots, X_m\}$ is aligned. For k-mer sequences with

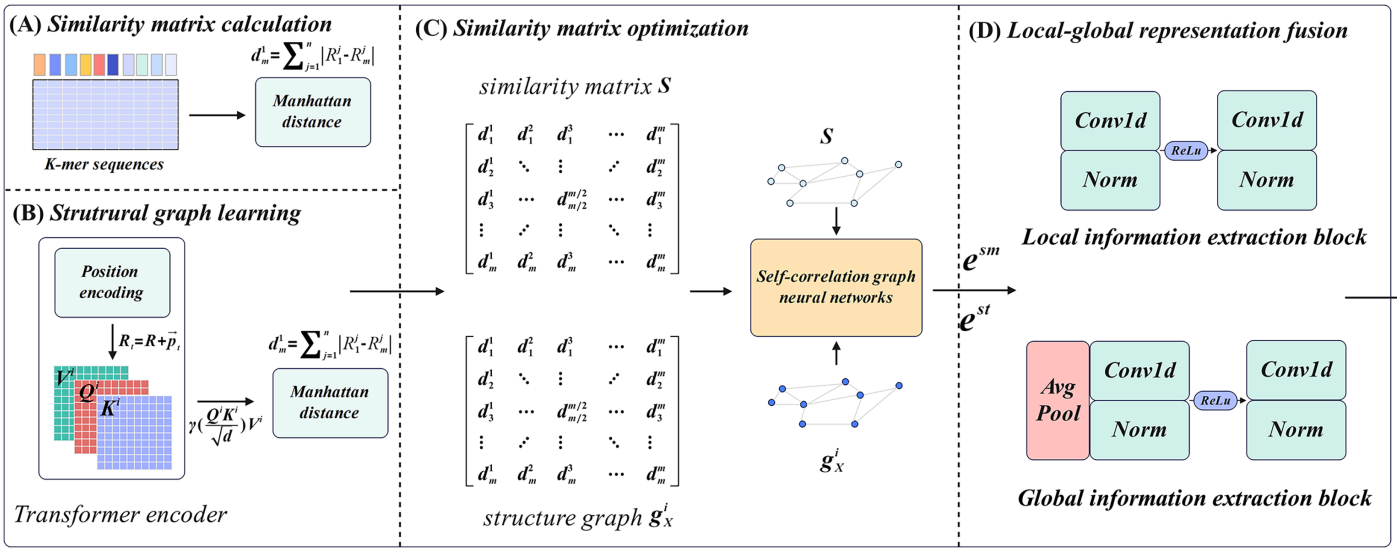

Figure 1 **The framework of M6ATMR.** (A) The details of similarity matrix calculation process. (B) The details of structural graph learning process with the Transformer encoder. (C) The similarity matrix optimization process with the self-correlation graph neural networks. (D) The structure of the local-global representation fusion block.
Full-size DOI: 10.7717/peerj.15899/fig-1

fixed K value, at the offset position l $(0 \leq l \leq n-k)$, we count the occurrence frequency of different substrings in k-mer sequence with length K starting from the offset position l:

$$R_X = Nom\left(\oplus_{i=l}^{n-k} fre\left(M_i\right)\right) \tag{1}$$

where $R_X \in \mathbb{R}^{n\times 1}$ denotes the k-mer representation of sequence X, and $fre(\cdot)$ represents the frequency of the k-mer substring M_i, and $\oplus$ represents the operation of concatenating all substring frequencies. $Nom(\cdot)$ represents a normalized operation.

After representing all RNA sequences using statistical k-mer frequency, we attempt to calculate the similarity between each sequence and construct an initial similarity matrix of RNA sequences. We construct matrices between k-mer representations for the following reasons: Firstly, earlier studies on the prediction of biomedical entity associations (*Shao et al., 2020*) have shown that similarity matrices can reflect the potential association between different entities, which provides strong support for the establishment of the association between different sequences of nucleotides. Secondly, constructing a graph data structure based on the similarity matrix allows us to capture the relationships between multiple sequences, which creates conditions for further sequence extraction. Therefore, it is a reasonable choice to transform the sequence representing the problem into the optimization task of the similarity matrix. In this article, we choose to employ the Manhattan distance to further measure the degree of similarity between nucleotides on different sequences:

$$S = \begin{pmatrix} d_1^m & \cdots & d_1^m \\ \vdots & \ddots & \vdots \\ d_m^1 & \cdots & d_m^m \end{pmatrix} \tag{2}$$

where, in the similar matrix $S \in \mathbb{R}^{m \times m}$, each element represents the Manhattan distance between k-mer sequences of corresponding positions. Taking d_1^m as an example, its value is the Manhattan distance between sequence representation R_1 and R_m:

$$d_m^1 = \sum_{j=1}^{n} \left| R_1^j - R_m^j \right| \tag{3}$$

where, the R_1^j, R_m^j represent the *jth* k-mer value in sequence state R_1 and R_m respectively, and $|\cdot|$ denotes the operation of taking the absolute value.

Structure graph learning

When obtaining the initial similarity matrices of the RNA sequences, we use a Transformer encoder to capture the structural information and learn the structural graph of RNA sequences. To this end, we apply the encoder part of the Transformer to further process the k-mer sequence. For a given Transformer encoder, there are two basic components: the position encoding block, and the self-attention mechanism. In addition, in order to further explore the structural relationship between RNA sequences, we also calculate the Manhattan distance between vector representations of the output of the Transformer encoding block, and strictly constrain the value of the structure matrix within the set of $\{0, 1\}$. The details are as follows.

Position encoding of Transformer is a functional encoder, that is, position vectors are calculated for each element in the sequence:

$$\overrightarrow{p_t}^{(i)} = f(t)^{(i)} := \begin{cases} \sin(w_i \cdot t) & i = 2k \\ \cos(w_i \cdot t) & i = 2k+1 \end{cases} \tag{4}$$

where w_i denotes the frequency, which is calculated as follows:

$$w_i = \frac{1}{10000^{2i/d}} \tag{5}$$

where d is the output dimension of the neural network. It is worth noting that the length of this position vector is equal to the length of k-mer representations. Thus, the RNA sequences are represented by k-mer and position encoding:

$$R_t = R + \overrightarrow{p_t} \tag{6}$$

where, $\overrightarrow{p_t}$ denotes the position vectors of all elements in k-mer representations. After calculating the position vectors, the encoder further optimizes the representation R_t through the self-attention mechanism:

$$\begin{cases} Q^i = w^q X^i \\ K^i = w^k X^i \\ V^i = w^v X^i \\ g_X^i = \gamma\left(\dfrac{Q^i (K^i)^T}{\sqrt{d}}\right) V^i \end{cases} \tag{7}$$

where, Q^i, K^i and V^i are the query matrix, key matrix and value matrix respectively. γ denotes the *softmax* activation function. Through the self-attention mechanism, the

Encoder can reveal the potential associations within the sequence and further excavate the structural associations of nucleotides.

In this regard, we believe that it is a valid way to describe the structural association of RNA sequences through internal relationships captured by the self-attention mechanism. To this end, we also measure this structural associations by the same method in section 2.3 and construct the structure graph G_{st} for sequences. Since the model samples the same data and distance formulas during the calculation process, the structure graph has a potential correlation with the similarity matrix, which indicates that it is reasonable to optimize the similarity matrix further through G_{st}. In addition, we ensure that the values of elements in G_{st} are strictly constrained in the set of $\{0, 1\}$ by threshold filtering, as shown below:

$$G_{st}^{(i,j)} = \begin{cases} 1 & if\ G(i,j) > 0.5 \\ 0 & otherwise \end{cases} \tag{8}$$

For each element $G(i, j)$ in $G_{st}^{(i,j)}$, updating the value to 0 if $G(i, j)$ is less than 0.5, otherwise updating the value to 1.

Similarity matrix optimization

To optimize the sequence similarity matrix and update the sequence structure graph, we design a self-correlation graph neural network that does not depend on prior node representations. In traditional graph neural networks, the embedded learning process relies on the existing representations of the nodes or edges in the graph. These prior representations serve as the starting point for the learning algorithm to update and refine the embeddings based on the graph structure. For instance, in some studies related to drug-drug association prediction, the SMILES of drugs are often applied as the prior information to serve as the initial input to the graph neural network. However, for similar matrix and structure graphs of sequences, the initial information of nodes is difficult to be obtained. In addition, both the similarity matrix and the structure graph describe the self-correlation property of the sequence, which makes the introduction of additional prior information may lead to misleading information. Therefore, inspired by the self-attention mechanism, we generate the learnable initial node representations based on the input matrix information. The initial representations are constantly updated and optimized during the process of graph convolution and participate in the optimization of the final embeddings. Taking the processing of similarity matrix as an example, we describe in detail the learning process of representation of similarity graph.

For a given similarity matrix $S \in \mathbb{R}^{m \times m}$, we define the node initial representation matrix as $Er \in \mathbb{R}^{m \times m}$. Each value in the matrix is determined and iteratively optimized by the network. S and Er are input into the three-layer self-correlation graph neural networks to get the embeddings related to the similarity matrix S:

$$\begin{cases} R_s^{(i+1)} = R_s^{(i)} + \alpha^{(i+1)} Gcov(Er, S)^{(i+1)} \\ \alpha^{(i+1)} = \dfrac{1}{I+1}(i+1) \end{cases} \tag{9}$$

where, α is the scaling superparameter to prevent the elements in the similarity matrix representation from becoming infinitesimal during graph convolution, and $I = 3$ denotes the number of convolution layers. $Gcov(\cdot)$ represents the convolution process. The hidden layer representation of layer $(i+1)$ and the representation of layer (i) satisfies the following equation:

$$H^{(l+1)} = \sigma(S^{-\frac{1}{2}}\left(diag(De_s) - \frac{1}{2}\left(S + S^T\right)\right)S^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{10}$$

where De_s denotes the degree matrix of the similar matrix S, and $diag(\cdot)$ denotes the diagonalization operation. W is the learnable weight. The representation of the hidden layer is initialized to Er, that is, $H^{(0)} = Er$. Similarly, sequence structure graphs are fed into the self-correlation neural networks in the same way. The network further employs learnable initial representations to mine the self-correlation of sequences, which also ensures consistency between representations learned from similarity matrices and that learned from structure graphs.

In addition, another task of the networks is to get better RNA representations by optimizing the similarity matrix. Hence, we apply the reconstructed similarity matrix and sequence structure graph to calculate the loss of the networks:

$$\mathcal{L}\left(\hat{S}, S, \widehat{G_{st}}, G_{st}\right) = BCELoss(\hat{S}, \mathrm{sigm}\left(S{*}S^T\right) + BCELoss(\widehat{G_{st}}, \mathrm{sigm}\left(G_{st}{*}{G_{st}}^T\right) \tag{11}$$

where $BCELoss(\cdot)$ denotes the binary cross-entropy loss. In the optimization process, we employ the difference between the reconstructed element values and the original element values to measure the performance of the representations.

Local-global representation fusion

In order to obtain comprehensive knowledge of RNA sequences, we propose a local-global strategy for fusing learned embedded representations. We process two kinds of representations of the same sequence respectively from the local and global perspectives of embedding representations, and further determine the weight relationship between embedding representations from similarity matrices and embedding representations from sequence structure graphs. Specifically, we design a local information extraction block and a global information extraction block respectively to calculate the weights of the two types of embedded representations.

For a given similarity matrix embedding representation e^{sm} and structure graph embedding representation e^{st}, we first treat them as residue sequences and weight them to obtain an overall representation e^a:

$$e^a = e^{sm} + e^{st} \tag{12}$$

We then enter e^a into the local and global information extraction blocks. For the local and global information extraction block, the extraction process is described as follows:

$$\begin{cases} e_{lo}^{a} = f(\vartheta(f'(e^{a}))) \\ e_{gl}^{a} = f(\vartheta(f'(\delta(e^{a})))) \end{cases} \quad (13)$$

where e_{lo}^{a} and e_{gl}^{a} denote the output of the local extraction block and that of the global extraction block respectively. f and f' represent a one-dimensional convolution layer containing normalized functions respectively, and ϑ represents the *ReLu* function. For the global information extraction block, we add the global average pooling layer δ on the basis of the local information extraction block. After that, we further calculate the weight difference w_f between the two representations:

$$w_f = sigm\left(e_{lo}^{a} + e_{gl}^{a}\right) \quad (14)$$

We utilize this weight difference to further integrate the two types of embedded representations to obtain the similar-structural representation e^{ss}:

$$e^{ss} = w_f * e^{sm} + \left(1 - w_f\right) * e^{st} \quad (15)$$

In addition, we consider the indispensable role of the k-mer representations of the sequences for the recognition of m6A modification sites, and further integrate these representations with the similar-structural representations to obtain the final embedding representations e^{fi}:

$$e^{fi} = w_{f'} * e^{ss} + \left(1 - w_{f'}\right) * e^{km} \quad (16)$$

where, e^{km} denotes the k-mer representations of the sequences, and $w_{f'}$ is the weight difference between e^{km} and e^{ss}.

Experiments setting

For a binary classification problem, its prediction states can be divided into the four categories: true positive (TP), false positive (FP), true negative (TN), false negative (FN). Thus, we select some predictive indicators to evaluate the prediction effect, and the calculation processes of these indicators are as below. Moreover, we obtain the area under the precision-recall curve (AUPR) and area under the receiver-operating characteristic curve (AUC) for evaluating our model.

$$Accuracy\ (Acc) = \frac{TP + TN}{TP + TN + FP + FN} \quad (17)$$

$$F1 = \frac{prec \times Sn}{prec + Sn} \quad (18)$$

$$Precision\ (Prec) = \frac{TP}{TP + FN} \quad (19)$$

$$Sensitivity\ (Sn) = \frac{TP}{TP + FN} \tag{20}$$

$$Specificity\ (Sp) = \frac{TN}{TN + FP} \tag{21}$$

$$Matthews\ Correlation\ Coefficient\ (MCC) = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{22}$$

In our study, we set the learning rate to 0.0001 and set the number of layers as three in the self-correlation graph neural network. Considering the time and complexity of training, we reduced the number of heads from eight to six in the Transformer encoder. We set the embedding size of the encoder to 32, the hidden dimension of the feed-forward layer to 128, and the number of encoder blocks to six.

RESULTS

Performance on A101 datasets

We conduct a rigorous 10-fold cross validation to evaluate the performance of our proposed model on the A101 dataset. The dataset is systematically partitioned into ten subsets of equal size, ensuring non-overlapping test sets in each fold. For each fold, we utilize nine subsets for training and one for testing. During evaluation, we consider six essential indicators: AUPR, AUC, Acc, F1 score, Prec, and Sen. The results of the 10-fold cross validation are presented through PR curves and ROC curves in Fig. 2 and summarized in Table 1. Across the 10 folds, the AUC curve demonstrates remarkable stability, with the maximum AUC reaching 93.87% and the minimum at 89.79%. Similarly, the PR curve displays consistent performance, with the highest AUPR at 93.17% and the lowest at 87.66%. Table 1 reveals outstanding performance in various evaluation metrics. The Acc achieves an impressive 84.43%, signifying a high correct identification rate for both TN and TP samples. Additionally, the MCC attains a value of 83.72%, reflecting the overall strength of our model. Furthermore, the values of other metrics, including F1 score, Prec, and Sen, surpass 80%, indicating the reliability of our model. These experimental findings substantiate the robustness and efficacy of our proposed model in identifying m6A modification sites.

Performance comparison of Classifiers

In some studies, the downstream classifiers have shown to significantly influence the classification performance of RNA sequence representations generated by models, potentially leading to model instability. To demonstrate the stability and efficacy of our proposed model, we conduct a comparison experiments using five additional classifiers: logistic regression, eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (lightgbm), CatBoost, and support vector machines (SVM). Logistic regression employs maximum likelihood estimation to predict model parameters, yielding binary results by minimizing cross-entropy loss during data training. XGBoost, an integrated

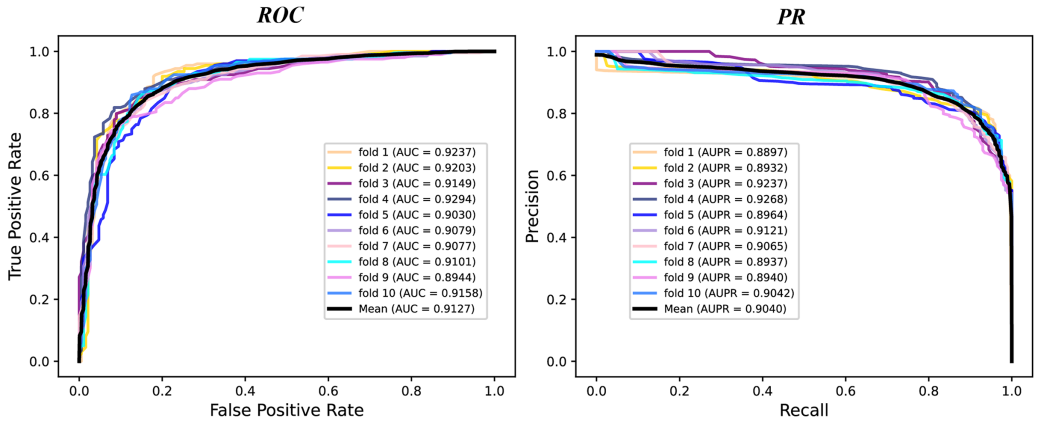

Figure 2 **The ROC curves and PR curves of M6ATMR on A101 dataset under 10-fold cross-validation.** Full-size DOI: 10.7717/peerj.15899/fig-2

Table 1 **The value of some indicators in each fold.**

Fold	MCC	Acc	Sn	Sp	Prec	F1	AUC	AUPR
0	0.7068	0.8535	0.8693	0.8368	0.8480	0.8586	0.9237	0.8897
1	0.7068	0.8535	0.8693	0.8368	0.8480	0.8586	0.9203	0.8932
2	0.6913	0.8458	0.8543	0.8368	0.8458	0.8500	0.9149	0.9237
3	0.7068	0.8535	0.8643	0.8421	0.8515	0.8579	0.9294	0.9268
4	0.6402	0.8201	0.8141	0.8263	0.8308	0.8223	0.9030	0.8964
5	0.6965	0.8483	0.8543	0.8421	0.8500	0.8521	0.9079	0.9121
6	0.6711	0.8355	0.8291	0.8421	0.8462	0.8376	0.9077	0.9065
7	0.7070	0.8535	0.8492	0.8579	0.8622	0.8557	0.9101	0.8937
8	0.6297	0.8149	0.8150	0.8148	0.8232	0.8191	0.8944	0.8940
9	0.7111	0.8557	0.8693	0.8413	0.8522	0.8607	0.9158	0.9042
Mean	0.6867	0.8434	0.8488	0.8377	0.8458	0.8473	0.9127	0.9040

classification algorithm, employs multiple simple base learners to iteratively train input data, continually reducing the discrepancy between model and input values. In contrast, lightgbm stands out with its advantage of low memory usage and faster training speed. CatBoost is designed to extract the most information from given data and is particularly effective for small machine-learning datasets. SVM, as a binary classification model, aims to find an optimal hyperplane for sample segmentation. For evaluation, we utilize 10-fold cross validation on the A101 dataset, as mentioned in "Performance on A101 datasets". The same set of indicators, AUPR, AUC, Acc, F1 score, Prec, and Sen are employed for assessment. The experimental results, presented in Fig. 3, Fig. 4, and Table 2, indicate that when RF is used as the downstream classifier, the model achieves the highest performance, with AUC at 91.27% and AUPR at 90.40%. While XGBoost exhibits relatively inferior performance compared to logistic regression and RF, the overall classification effect of all three classifiers remains relatively favorable. The difference in AUC values between XGBoost and RF is 7.84%. These findings support the notion that the RNA sequence

Figure 3 **The comparison results of different classifiers.** Full-size DOI: 10.7717/peerj.15899/fig-3

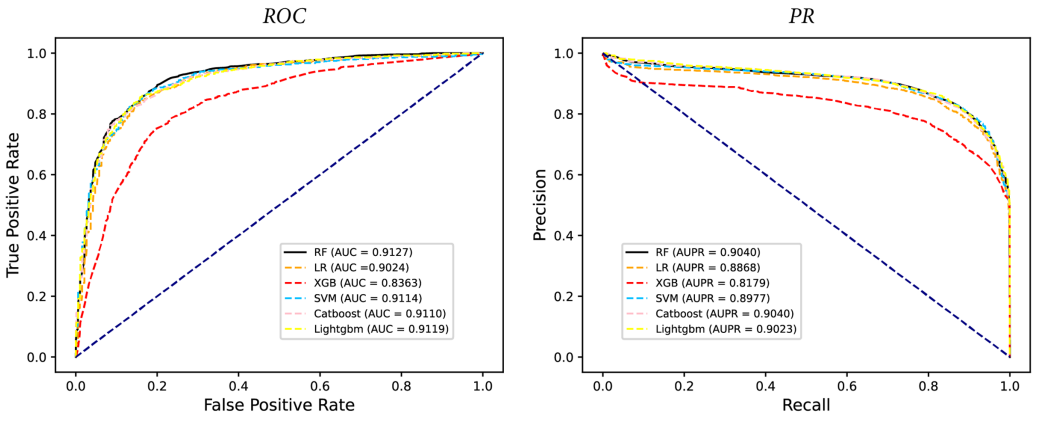

Figure 4 **The PR curves of M6ATMR with different classifiers.**
Full-size DOI: 10.7717/peerj.15899/fig-4

Table 2 **The AUC value of some indicators in each fold based on different classifiers.**

Fold	RF	LR	XGBoost	SVM	CatBoost	Lightgbm
Fold1	0.9237	0.9052	0.8454	0.9077	0.8951	0.9111
Fold2	0.9203	0.8932	0.8319	0.9085	0.9484	0.9305
Fold3	0.9149	0.9148	0.8125	0.8856	0.8928	0.8985
Fold4	0.9294	0.8806	0.8714	0.9270	0.9078	0.9080
Fold5	0.9030	0.9282	0.7987	0.9216	0.9089	0.8996
Fold6	0.90793	0.8855	0.8208	0.9033	0.9040	0.8752
Fold7	0.9077	0.9025	0.7991	0.9141	0.9284	0.9255
Fold8	0.9101	0.9022	0.8470	0.9108	0.8961	0.9412
Fold9	0.8944	0.9030	0.8821	0.9138	0.9195	0.9248
Fold10	0.9158	0.9087	0.8537	0.9290	0.9094	0.9052

representations learned by our model are not easily influenced by the choice of downstream classifiers, indicating the model's stability and effectiveness in feature extraction and m6A modification site identification.

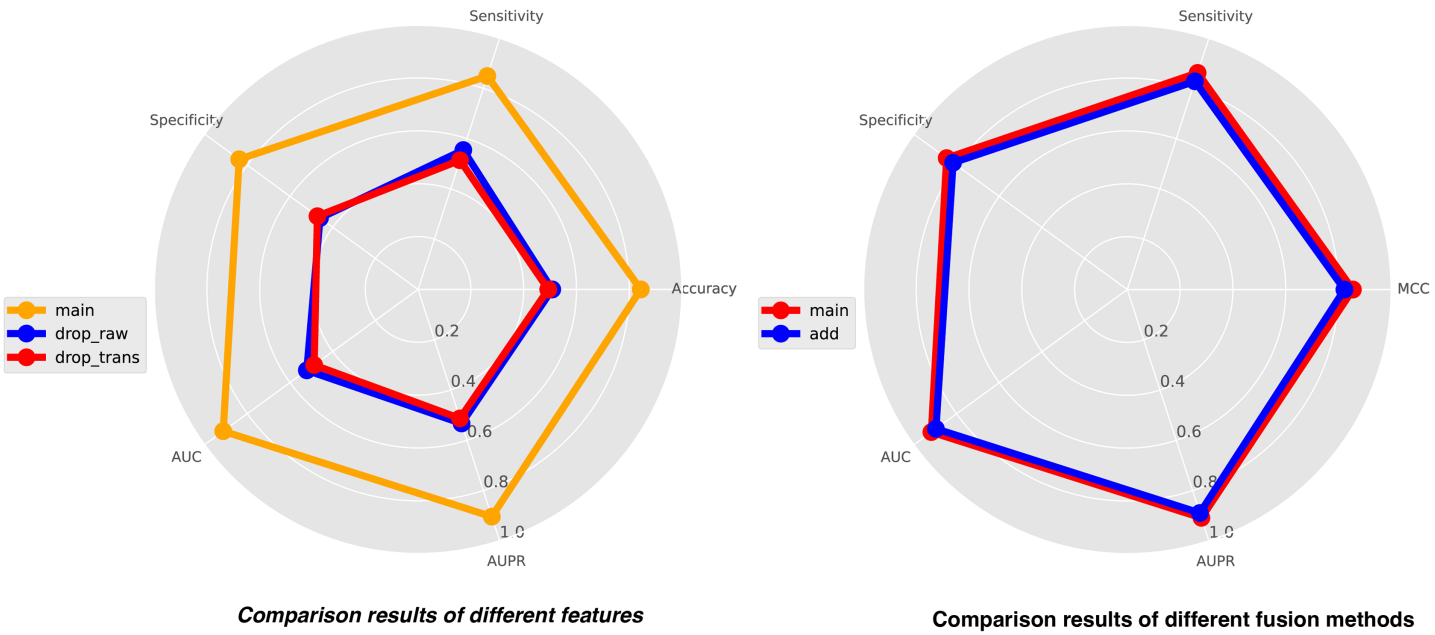

Figure 5 **The comparison results of different features (left) and different fusion methods (right).** Full-size DOI: 10.7717/peerj.15899/fig-5

Features selection

In this study, we integrate three types of features for comprehensive analysis, including feature representations from similarity matrices, feature representations from sequence structure diagrams, and k-mer sequence representations. K-mer representations are pre-coded representations based on the frequency count of sequence k-mer substrings, while the other two features are dynamically learned through neural networks. To achieve a holistic understanding of the sequences, we perform local-global integration of these features. To illustrate the importance of combining these three features, we construct two additional features, drop-raw and drop-trans, to validate our model's performance. Drop-raw represents features without similarity matrix information, and drop-trans represents features without sequence structure graph information. The comparison results of the three types of features are presented in Fig. 5. Our model demonstrates the best recognition performance when all features are utilized. Drop-raw performs slightly better than drop-trans in terms of AUC and AUPR, suggesting that similarity matrices exert a stronger influence on the model compared to sequence structure graphs. Overall, all three types of features are essential and significantly enhance the effectiveness of site recognition. The local-global fusion block is a crucial component of our model, which integrates multiple features from both local and global perspectives by combining learned similarity matrix features with sequence-structure graph features. To demonstrate the necessity of this fusion block, we design another strategy using only weighted fusion, and the results are also presented in Fig. 5. The outcomes reveal that the recognition performance of the model is inferior when only weighted fusion is employed compared to the local-global fusion block. This underscores the significance of the local-global fusion block in achieving

Table 3 The comparison results of different recognition methods. N.A. denotes the value of the indicator is not provided by corresponding studies.

Models	Acc	MCC	Sn	Sp
M6ATMR	0.8434	0.6867	0.8488	0.83.77
M6AMRFS	0.8105	0.6210	0.8067	0.81.43
BERMP	N.A.	0.7260	0.8230	0.90.00
RFAthM6A	N.A.	0.7255	0.8222	0.90.00
BERT-m7G	0.8213	0.7023	0.8124	0.7985
Le & Ho (2022)	0.7325	0.5254	0.7234	0.6638

superior performance and highlighting its role in effectively capturing comprehensive sequence information.

Performance comparison of predictors

In this study, we conduct a rigorous 10-fold cross validation to compare our model with five other existing methods on the A101 dataset. The compared methods include M6AMRFS, BERMP, RFAthM6A, BERT-m7G (*Zhang et al., 2021*), and the model designed by *Le & Ho (2022)*. M6AMRFS encodes RNA sequences using two feature descriptors, dinucleotide binary coding, and local site-specific dinucleotide frequency. It enhances feature representation through the F-score algorithm combined with sequence forward search (SFS) and employs XGBoost as the downstream classifier. BERMP utilizes GRU to represent RNA sequences and adopts an end-to-end training process for site recognition. RFAthM6A attempts to classify various types of features derived from RNA sequences using machine learning methods. Our model, M6ATMR, adopts the transformer encoder to extract sequence representations and uses a stacking ensemble classifier for predicting m6A sites. We also consider two other transformer-based models, BERT-m7G, and the model designed by *Le & Ho (2022)*. In their approaches, BERT-m7G uses bidirectional encoder representations from transformers (BERT) to extract sequence representations, while *Le & Ho (2022)* use a pre-trained transformer to explore features and a convolutional neural network for further feature extraction. The comparison results are presented in Table 3. Our model achieves an Acc value of 84.42% and a MCC value of 83.72%. These indicators demonstrate that our model outperforms the other five methods in most aspects. Compared to the other models, our approach exhibits a remarkable improvement, with a maximum of 11.09% higher accuracy and 16.13% higher MCC value. The model designed by *Le & Ho (2022)* has the lowest MCC value among all methods, while BERMP and RFAthM6A show similar performance across various indicators. However, we note that the specificity (Sp) value of our method is slightly lower than that of BERMP and RFAthM6A, indicating a minor deficiency in predicting true negative samples. Nevertheless, overall, the experimental results clearly demonstrate the effectiveness of our model, which stands out as a superior approach for m6A site prediction on RNA sequences.

DISCUSSION AND CONCLUSION

In this article, we commence by reviewing classical methods for identifying RNA m6A modification sites and presenting our own perspectives. Subsequently, we analyze the limitations of these methods, leading us to propose a novel sequence-dependent-only RNA m6A modification site recognition method, named M6ATMR. M6ATMR utilizes the Transformer to guide the reconstruction of similarity matrices for each RNA sequence, thereby optimizing the feature representation of RNA sequences. Comparative analysis with other recognition methods reveals that M6ATMR demonstrates superior predictive performance, as evidenced by improved metrics. Comprehensive experiments further attest to the accuracy and robustness of our model. Additionally, we delve into several critical aspects. First, computing the similarity matrix and optimizing feature generation proves effective in enhancing the recognition performance of RNA m6A modification sites. Second, cooperative updating of similarity matrices and sequence structure graphs in the sequence representation of the same RNA sequence facilitates the retention of richer nucleotide distribution information. Third, the deep fusion of multiple features from both local and global perspectives results in a comprehensive understanding of RNA sequences.

However, there remain certain limitations in our study that warrant attention. First, the restriction of RNA sequence length necessitates the selection of the A101 dataset for model verification, rendering our approach less adept at handling short RNA sequences. Second, our current model primarily focuses on nucleotide distribution information, with limited exploration of other sequence properties. Future work will address these issues and explore the application of our model to the identification of other modification types, such as M1A, and modifications on DNA sequences.

ACKNOWLEDGEMENTS

The authors thank the participants for their cooperation in the study.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests

The authors are employed by Changjiang water resources and hydropower development group.

Author Contributions

- Shuang Xiang conceived and designed the experiments, performed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Te Zhang performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.

- Minghao Wu conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Raw data are available in the Supplemental Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.15899#supplemental-information.

REFERENCES

Breiman L. 2001. Random forests. *Machine Learning* **45(1)**:5–32 DOI 10.1023/A:1010933404324.

Chen W, Ding H, Zhou X, Lin H, Chou K-C. 2018. iRNA (m6A)-PseDNC: identifying N6-methyladenosine sites using pseudo dinucleotide composition. *Analytical Biochemistry* **561**:59–65 DOI 10.1016/j.ab.2018.09.002.

Chen W, Feng P, Ding H, Lin H. 2016. Identifying N 6-methyladenosine sites in the *Arabidopsis thaliana* transcriptome. *Molecular Genetics and Genomics* **291(6)**:2225–2229 DOI 10.1007/s00438-016-1243-7.

Chen W, Tran H, Liang Z, Lin H, Zhang L. 2015. Identification and analysis of the N6-methyladenosine in the *Saccharomyces cerevisiae* transcriptome. *Scientific Reports* **5**:13859 DOI 10.1038/srep13859.

Di Giallonardo F, Schlub TE, Shi M, Holmes EC. 2017. Dinucleotide composition in animal RNA viruses is shaped more by virus family than by host species. *Journal of virology* **91(8)**:e02381-16 DOI 10.1128/JVI.02381-16.

Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M. 2012. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **485(7397)**:201–206 DOI 10.1038/nature11112.

Fustin J-M, Doi M, Yamaguchi Y, Hida H, Nishimura S, Yoshida M, Isagawa T, Morioka MS, Kakeya H, Manabe I, Okamura H. 2013. RNA-methylation-dependent RNA processing controls the speed of the circadian clock. *Cell* **155(4)**:793–806 DOI 10.1016/j.cell.2013.10.026.

Geula S, Moshitch-Moshkovitz S, Dominissini D, Mansour AA, Kol N, Salmon-Divon M, Hershkovitz V, Peer E, Mor N, Manor YS. 2015. m6A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation. *Science* **347(6225)**:1002–1006 DOI 10.1126/science.1261417.

Huang Y, He N, Chen Y, Chen Z, Li L. 2018. BERMP: a cross-species classifier for predicting m6A sites by integrating a deep learning algorithm and a random forest approach. *International Journal of Biological Sciences* **14(12)**:1669–1677 DOI 10.7150/ijbs.27819.

Ke S, Alemu EA, Mertens C, Gantman EC, Fak JJ, Mele A, Haripal B, Zucker-Scharff I, Moore MJ, Park CY. 2015. A majority of m6A residues are in the last exons, allowing the potential for 3′ UTR regulation. *Genes & Development* **29(19)**:2037–2053 DOI 10.1101/gad.269415.115.

Le NQK, Ho Q-T. 2022. Deep transformers and convolutional neural network in identifying DNA N6-methyladenine sites in cross-species genomes. *Methods* **204(1)**:199–206 DOI 10.1016/j.ymeth.2021.12.004.

Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR. 2012. Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell* **149(7)**:1635–1646 DOI 10.1016/j.cell.2012.05.003.

Nazari I, Tahir M, Tayara H, Chong KT. 2019. iN6-Methyl (5-step): Identifying RNA N6-methyladenosine sites using deep learning mode via Chou's 5-step rules and Chou's general PseKNC. *Chemometrics and Intelligent Laboratory Systems* **193(3)**:103811 DOI 10.1016/j.chemolab.2019.103811.

Qiang X, Chen H, Ye X, Su R, Wei L. 2018. M6AMRFS: robust prediction of N6-methyladenosine sites with sequence-based features in multiple species. *Frontiers in Genetics* **9**:495 DOI 10.3389/fgene.2018.00495.

Qiu W-R, Jiang S-Y, Sun B-Q, Xiao X, Cheng X, Chou K-C. 2017. iRNA-2methyl: identify RNA 2'-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier. *Medicinal Chemistry* **13(8)**:734–743 DOI 10.2174/1573406413666170623082245.

Rehman MU, Hong KJ, Tayara H, to Chong K. 2021. convolution neural tool for RNA N6-Methyladenosine site identification in different species. *IEEE Access* **9**:17779–17786 DOI 10.1109/ACCESS.2021.3054361.

Shao K, Zhang Z, He S, Bo X. 2020. DTIGCCN: prediction of drug-target interactions based on GCN and CNN. In: *Paper presented at: 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*. Piscataway: IEEE.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 2017. Attention is all you need. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.

Wan Y, Tang K, Zhang D, Xie S, Zhu X, Wang Z, Lang Z. 2015. Transcriptome-wide high-throughput deep m6A-seq reveals unique differential m6A methylation patterns between three organs in *Arabidopsis thaliana*. *Genome Biology* **16(1)**:1–26 DOI 10.1186/s13059-015-0839-2.

Wang Y, Guo R, Huang L, Yang S, Hu X, He K. 2021. A predictor for n6-methyladenosine sites identification utilizing sequence characteristics and graph embedding-based geometrical information. *Frontiers in Genetics* **12**:670852 DOI 10.3389/fgene.2021.670852.

Wang Y, Li Y, Toth JI, Petroski MD, Zhang Z, Zhao JC. 2014. N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nature Cell Biology* **16(2)**:191–198 DOI 10.1038/ncb2902.

Wang Y, Li Y, Yue M, Wang J, Kumar S, Wechsler-Reya RJ, Zhang Z, Ogawa Y, Kellis M, Duester G. 2018. N6-methyladenosine RNA modification regulates embryonic neural stem cell self-renewal through histone modifications. *Nature Neuroscience* **21(2)**:195–206 DOI 10.1038/s41593-017-0057-1.

Wang J, Wang L. 2020. Deep analysis of RNA N6-adenosine methylation (m6A) patterns in human cells. *NAR Genomics and Bioinformatics* **2(1)**:lqaa007 DOI 10.1093/nargab/lqaa007.

Wang X, Yan R. 2018. RFAthM6A: a new tool for predicting m6A sites in *Arabidopsis thaliana*. *Plant Molecular Biology* **96(3)**:327–337 DOI 10.1007/s11103-018-0698-9.

Xiang S, Yan Z, Liu K, Zhang Y, Sun Z. 2016. AthMethPre: a web server for the prediction and query of mRNA m 6 A sites in *Arabidopsis thaliana*. *Molecular BioSystems* **12(11)**:3333–3337 DOI 10.1039/C6MB00536E.

Xing P, Su R, Guo F, Wei L. 2017. Identifying N6-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine. *Scientific reports* **7**:46757 DOI 10.1038/srep46757.

Yang H, Lv H, Ding H, Chen W, Lin H. 2018. a sequence-based predictor for identifying 2′-O-methylation sites in Homo sapiens. *Journal of computational biology* **25(11)**:1266–1277 DOI 10.1089/cmb.2018.0004.

Zhang L, Dong B, Teng Z, Zhang Y, Juan L. 2020. Identification of human enzymes using amino acid composition and the composition of-spaced amino acid pairs. *BioMed Research International* **2020(1)**:1–11 DOI 10.1155/2020/9235920.

Zhang L, Qin X, Liu M, Liu G, Ren Y. 2021. BERT-m7G: a transformer architecture based on BERT and stacking ensemble to identify RNA N7-Methylguanosine sites from sequence information. *Computational and Mathematical Methods in Medicine* **2021**:7764764 DOI 10.1155/2021/7764764.

Zhou Y, Zeng P, Li Y-H, Zhang Z, Cui Q. 2016. SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. *Nucleic Acids Research* **44(10)**:e91 DOI 10.1093/nar/gkw104.