BMC Bioinformatics

RESEARCH Open Access

MSCAN: multi-scale self- and cross-attention network for RNA methylation site prediction

Honglei Wang[1,3], Tao Huang[1], Dong Wang[2], Wenliang Zeng[1], Yanjing Sun[1*] and Lin Zhang[1*]

*Correspondence:
yjsun@cumt.edu.cn;
lin.zhang@cumt.edu.cn

[1] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China
[2] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China
[3] School of Information Engineering, Xuzhou College of Industrial Technology, Xuzhou 221400, China

Abstract

Background: Epi-transcriptome regulation through post-transcriptional RNA modifications is essential for all RNA types. Precise recognition of RNA modifications is critical for understanding their functions and regulatory mechanisms. However, wet experimental methods are often costly and time-consuming, limiting their wide range of applications. Therefore, recent research has focused on developing computational methods, particularly deep learning (DL). Bidirectional long short-term memory (BiLSTM), convolutional neural network (CNN), and the transformer have demonstrated achievements in modification site prediction. However, BiLSTM cannot achieve parallel computation, leading to a long training time, CNN cannot learn the dependencies of the long distance of the sequence, and the Transformer lacks information interaction with sequences at different scales. This insight underscores the necessity for continued research and development in natural language processing (NLP) and DL to devise an enhanced prediction framework that can effectively address the challenges presented.

Results: This study presents a multi-scale self- and cross-attention network (MSCAN) to identify the RNA methylation site using an NLP and DL way. Experiment results on twelve RNA modification sites (m^6A, m^1A, m^5C, m^5U, m^6Am, m^7G, Ψ, I, Am, Cm, Gm, and Um) reveal that the area under the receiver operating characteristic of MSCAN obtains respectively 98.34%, 85.41%, 97.29%, 96.74%, 99.04%, 79.94%, 76.22%, 65.69%, 92.92%, 92.03%, 95.77%, 89.66%, which is better than the state-of-the-art prediction model. This indicates that the model has strong generalization capabilities. Furthermore, MSCAN reveals a strong association among different types of RNA modifications from an experimental perspective. A user-friendly web server for predicting twelve widely occurring human RNA modification sites (m^6A, m^1A, m^5C, m^5U, m^6Am, m^7G, Ψ, I, Am, Cm, Gm, and Um) is available at http://47.242.23.141/MSCAN/index.php.

Conclusions: A predictor framework has been developed through binary classification to predict RNA methylation sites.

Keywords: RNA methylation, Transformer, Predictor, Multi-scale, Self-attention, Cross-attention

Background

RNA modification plays a fundamental role in regulating RNA function [1] and has become a hotspot in epigenetics research [2]. Nearly 200 RNA modifications have been discovered, most of which are methylation modifications [3]. Common RNA methylation types include N^1-methyladenosine (m^1A), N^2-methylguanosine (m^2G), 5-methylcytosine (m^5C), 5-methyluridine (m^5U), 2′-*O*-methyladensine (Am), 2′-*O*-methylcytidine (Cm), 2′-*O*-methylguanosine (Gm), 2′-*O*-methyluridine (Um), Pseudouridine (Ψ), N^6-methyladenosine(m^6A), N^7-methylguanine (m^7G), inosine (I), and N6,2′-*O*-dimethyladenosine(m^6Am), etc. Among them, m^6A refers to methylation modification occurring at the nitrogen atom in position 6 of the RNA molecule adenine, which is the most abundant mRNA methylation, and is known to affect mRNA stability, splicing, and translation. In addition to m^6A, m^1A RNA methylation is a recently discovered one, which is evolutionarily conserved and ubiquitous in humans, rodents, and yeast. It can significantly enhance the protein translation of transcripts [4], block the Watson–Crick interface, and is essential for tRNA stability [5].

In the last decade, dozens of experimental methods have been developed to identify the precise location of methylation sites on RNA, such as miCLIP [6], m^1A-seq [7], PA-m6A-seq [8], m^1A-ID-seq [9], m5C-RIP [10], m^1A-MAP [11], and m^1A-IP-seq [12]. Despite their effectiveness, these experimental techniques are usually both time-consuming and costly, limiting their use in different biological contexts [4], and making them inadequate for large-scale genomic data [13]. Consequently, there is strong motivation to explore computational methods that can accurately and efficiently identify methylation sites based on sequence information alone.

As there are more vailable base-resolution datasets, researchers have designed some computational methods for RNA modification site prediction. These approaches formulate RNA methylation identification as a binary prediction task, and some machine learning models are trained to distinguish between truly methylated and non-methylated sites. These computational methods have been powerful additions for RNA methylation site prediction.

Traditional methods designed for sequence-based prediction usually first extract features based on human-understandable feature methods and then use a classifier to identify if the site is methylated based on the preceding extracted features. Specifically, RAMPred [14] adopts the support vector machine(SVM) to predict the m^1A modification site, extracting features based on nucleotide composition(NC) and nucleotide chemical properties(NCP). iRNA-3typeA [15] adopts SVM to predict m^1A, A-to-I, and m^6A modification sites, which extracts features based on accumulated nucleotide frequency(ANF) and NCP. iMRM [16] extracts features based on NCP, NC, Nucleotide Density(ND), Dinucleotide physicochemical properties(DPCP), and Dinucleotide Binary Encoding(DBE) and employs XGboost to predict m^1A, m^6A, m^5C, ψ, and A-to-I modification sites. The above sequence features are artificially extracted, and inevitably important features of the sequences are missed due to human cognitive limitations.

Analyzing biological sequences and interpreting biological information are the key challenges in achieving biological discovery. The application of natural language processing(NLP) to sequence analysis has attracted considerable attention in processing biological sequences [17]. As biological sequences can be considered sentences,

and k-mer subsequences are regarded as words [18, 19], NLP can be used to understand the structure and function encoded in these sequences [17]. Unlike traditional machine learning, deep learning (DL) methods follow an end-to-end design. Features are extracted directly based on the input sequence and the final labeling/prediction task. For example, EDLm6Apred [20] employs bidirectional long short-term memory (BiLSTM) to predict m^6A sites, extracting features based on Word2vec, RNA word embedding [21], and one-hot encoding [22, 23]. However, LSTM, BiLSTM, and RNN cannot achieve parallel computation, leading to a long training time.

CNN can achieve parallel computation and learn local dependencies. For instance, m6A-word2vec [24] adopts CNN to identify m^6A sites, extracting features based on Word2vec. Deeppromise [25] employs CNN to identify m^1A and m^6A sites, extracting features based on integrated enhanced nucleic acid composition (ENAC) [26], one-hot encoding, and RNA word embedding. However, These CNN structures only consider the contextual relationships of neighboring bases without considering the dependencies over long distances in the sequence. DeepM6ASeq [27] combines the advantages of CNN and BiLSTM by using two layers of CNN and one layer of BiLSTM to predict m^6A sites. This approach may extract redundant features that interfere with prediction performance [28]. The attention mechanism can quantify the degree of code-to-code dependency [29]. Therefore, the application of the attention mechanism can capture the focused codes that affect the classification results. Plant6mA [30] utilizes a Transformer encoder to determine whether the input sequence contains an m6A site. However, due to the unique feature representation of transformers, these networks are primarily employed at a single scale. Although a single-scale self-attentive mechanism can focus on essential features of sequence context, it lacks information interaction with sequences at different scales. It isn't easy to learn complex word context relationships.

At present, most prediction model studies focus only on a single methylation modification, and few share the same binary classification model framework to achieve different methylation modification predictions. Even fewer cross-modification validation studies have been performed with different methylation test sets and trained models. Accounting for potential interactions between various RNA modifications, it would be interesting to use the same model to conduct cross-modification validation studies across different methylation test sets.

We present the Multi-scale Self- and Cross-attention Network (MSCAN), a novel approach designed to identify RNA methylation sites, addressing the challenges associated with current methods. Our model supports identifying twelve RNA modification types, including m^6A, Ψ, m^1A, m^6Am, Am, Cm, m^7G, Gm, Um, I, m^5U, and m^5C.

The MSCAN employs a unique multi-scale approach for analyzing RNA sequences. Specifically, we extracted the input 41-nucleotides (nt) sample sequence into multiple smaller subsequences centered around the sequence midpoint. To ensure accurate identification of methylation sites, the MSCAN analyzes these smaller subsequences at two distinct scales: 21-nt and 31-nt. This multi-scale analysis allows for a more comprehensive understanding of the RNA sequence context, ultimately leading to improved prediction performance. Secondly, word2vec was used to encode the three sets of sequences. Third, the three sets of sequences add positional information due to the correlation between nucleotide positions in the sequence. Four, the three sets of sequences were

fed into the encoding module, which was constructed with a multi-scale self- and cross-attention network and a feed-forward network(FFN) to extract potential contributing features for methylation site prediction. Finally, methylation predicted probabilities were obtained through a linear layer and the sigmoid function. The findings demonstrated that the MSCAN model surpassed the performance of state-of-the-art methods, including m6A-word2vec, DeepM6ASeq, and Plant6mA in independent tests. A user-friendly web server for MSCAN is available at http://47.242.23.141/MSCAN/index.php.

Result

Evaluation metrics

In this study, we used eight common classification indicators to evaluate the prediction of the model, including Accuracy (Acc), Sensitivity (Sen), Precision (Pre), Matthews correlation coefficient (MCC), Specificity (Sp), and F1 score (F1). The formulas of these metrics are as follows:

$$Sensitivity,\ Recall = \frac{TP}{TP + FN} \tag{1}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$F1\,score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{6}$$

Here, the true positive, true negative, false positive, and false negative are represented as TP, TN, FP, and FN, respectively. Moreover, the area under the receiver operating characteristic (AUROC) and the area under the precision-recall curve (AUPRC) are used to visually evaluate the model's overall performance.

Results analysis

MSCAN completed model training and experimental parameter optimization based on the dataset of Chen et al. [25]. Subsequently, MSCAN completed the model's generalization ability evaluation based on the dataset of Song et al. [5]. Specifically, based on the dataset of Chen et al., this paper first compared the performance of MSCAN with different combinations of input sequences on the training data. Second, the performance of MSCAN with different feature encoding was compared. Third, we compared the performance of different MSCAN model variants. Fourth, the MSCAN was compared with

Table 1 Evaluation results of five-fold cross-validation of transformer based on the training data of Chen et al.

Length(nt)	AUROC	ACC	Sen	Precision	MCC	Spe	F-1	AUPRC
11	0.8955	93.55	44.26	**77.14**	55.44	**98.65**	56.25	0.6389
21	0.9213	**95.01**	59.81	74.42	64.11	98.16	66.32	0.7189
31	0.9361	93.55	64.80	66.94	62.31	96.61	65.85	0.7390
41	**0.9484**	93.48	**74.36**	61.27	63.97	95.37	67.18	**0.7469**
51	0.9401	94.63	61.54	74.23	64.73	97.89	67.29	0.7401
61	0.9430	94.56	67.27	67.89	64.61	97.07	67.58	0.7078
71	0.9249	93.64	67.21	65.60	62.89	96.36	66.40	0.7272
81	0.9440	94.94	70.00	66.04	**65.25**	97.01	**67.96**	0.7170
91	0.9327	94.93	60.19	73.86	64.01	98.08	66.33	0.7119
101	0.9230	93.33	70.94	61.03	62.16	95.53	65.61	0.7449

Bold indicates the best performance

state-of-the-art models based on the training data and the independent dataset of Chen et al. Fifth, the statistical significance of AUPRC values between the four models is compared. Sixth, MSCAN completed a generalization ability evaluation based on the dataset of Song et al., and MSCAN outperformed the state-of-the-art predictors for twelve modification sites. Finally, We designed a cross-modification validation experiment in which twelve models with different methylation types were compared for prediction performance based on twelve test sets, respectively. Our experiments were conducted with two Intel(R) 5218 CPUs, two RTX2080Ti GPUs, and Pytorch version 1.4.0+cu92.

Based on the training data of Chen et al., we first tried optimizing the input sequences' length according to AUPRC on the training data. Using the Word2vec embedding, we evaluated the Transformer model with 11-nt, 21-nt, 31-nt, 41-nt,51-nt,61-nt,71-nt,81-nt,91-nt,and 101-nt RNA sequences as the input on the five-fold cross-validation [5, 25]. As shown in Table 1, the input of the 11-nt sequence obtained the worst performance. The reason may be that too few bases in the 11-nt sequence affect feature extraction. The input of the 41-bp sequence obtained the best average performance of all the modifications, It may be worth mentioning that the 41-nt of the input sequence is also optimal for the XGboost and SVM method [14, 16], so we choose 21-nt, 31-nt, and 41-nt RNA sequences as input sequences to achieve different combinations of input sequences.

Table 2 Evaluation results of MSCAN on five-fold cross-validation with different input sequences based on the training data of Chen et al.

combinations of sequences(nt)	AUROC	ACC	Sen	Precision	MCC	Spe	F-1	AUPRC
21 + 31 + 41	0.9491	**95.24**	58.59	73.42	63.11	98.26	65.17	0.7695
21 + 41 + 31	**0.9618**	94.86	59.69	**83.70**	**68.11**	**98.72**	69.68	**0.7949**
31 + 21 + 41	0.9257	94.63	55.93	78.57	63.59	98.48	65.34	0.7419
31 + 41 + 21	0.9463	95.09	**68.47**	72.38	67.73	97.57	**70.37**	0.7632
41 + 21 + 31	0.9318	94.55	63.87	73.08	65.38	97.64	68.17	0.7617
41 + 31 + 21	0.9427	93.86	65.41	71.90	65.20	97.10	68.50	0.7642

Bold indicates the best performance

The combination of input sequences with different scale order is an important parameter that affects the performance of the training model. The performance of the MSCAN model with the different combinations of input sequences on the training data is shown in Table 2. MSCAN shows the best prediction performance when the combination is "21-nt + 41-nt + 31-nt". According to the MSCAN model design, "21-nt + 41-nt + 31-nt" input sequences are entered into the model to implement three attention mechanisms, including the self-attention calculation mechanism for the 21-nt sequence, and the cross-attention calculation mechanisms for both "21-nt + 41-nt" and "21-nt + 31-nt" combinatorial sequences.

Comparison analysis of different feature encoding methods

In this section, we evaluate the performance of three distinct feature encoding methods—Word2vec, One-hot, and ENAC—utilizing the same MSCAN model for predicting m^1A sites on the test data of Chen et al. The outcomes of this comparison are presented in Fig. 1 and Table 3, demonstrating that Word2vec consistently surpasses the other two encoding methods across all performance indices.

The superior performance of Word2vec can be attributed to the limitations of the One-hot and ENAC encoding methods. While One-hot encoding focuses on the local information of individual bases, and ENAC encoding considers both nucleic acid composition and position information, both methods neglect the semantic information inherent in the sequence context. In contrast, Word2vec prioritizes the contextual relationships between bases, resulting in a more effective representation of the sequence.

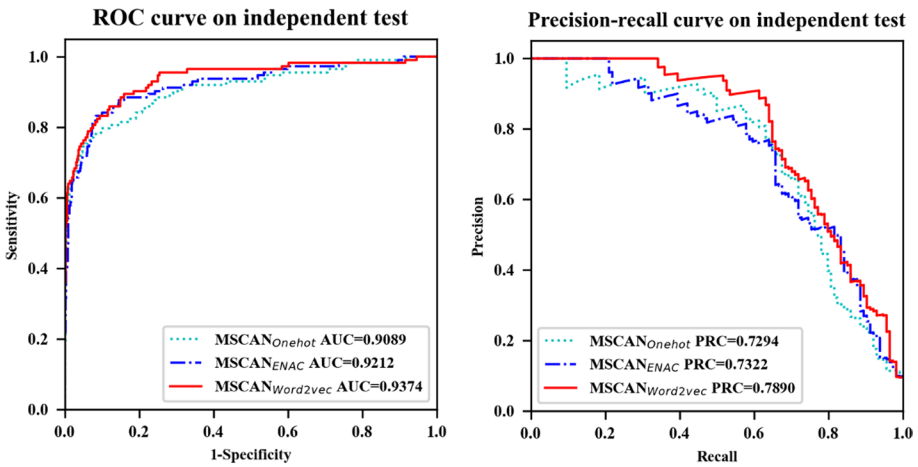


Fig. 1 Performance of the MSCAN model based on the different feature encoding

Table 3 MSCAN model evaluation results with different feature encodings based on the test data of Chen et al.

Encoding	AUROC	ACC	Sen	Precision	MCC	Spe	F-1	AUPRC
One-hot	0.9089	95.13	55.26	86.30	66.77	99.12	67.38	0.7294
ENAC	0.9212	94.81	55.26	81.82	64.71	98.77	65.97	0.7322
Word2vec	**0.9374**	**95.69**	**58.77**	**90.54**	**70.95**	**99.39**	**71.28**	**0.7890**

Bold Indicates the best performance

Our findings highlight the importance of selecting appropriate feature encoding methods for improved prediction accuracy, with Word2vec emerging as a particularly advantageous choice for the MSCAN model in the context of RNA methylation site prediction.

Comparison with different variants of the MSCAN model

We conducted ablation experiments to assess the contribution of key components within our proposed MSCAN model based on the test data of Chen et al. Utilizing Word2vec for RNA sequence encoding, we constructed four sub-networks: self- and cross-attention network (SCAN), self-attention network (SAN), multi-scale cross-attention network (MCAN), and cross-attention network (CAN). SCAN represents MSCAN with one cross-attention module removed, SAN is SCAN devoid of cross-attention, MCAN is MSCAN without self-attention, and CAN is MCAN with one cross-attention module removed. The outcomes of these experiments are depicted in Fig. 2 and summarized in Table 4.

SAN serves as the baseline model in this comparison. Upon the integration of cross-attention modules, the area under the precision-recall curve (AUPRC) for SCAN and MSCAN models increased by 0.09% and 2.86%, respectively. These results highlight the importance of incorporating cross-attention mechanisms within the MSCAN model for improved performance in predicting RNA methylation sites. Consequently,

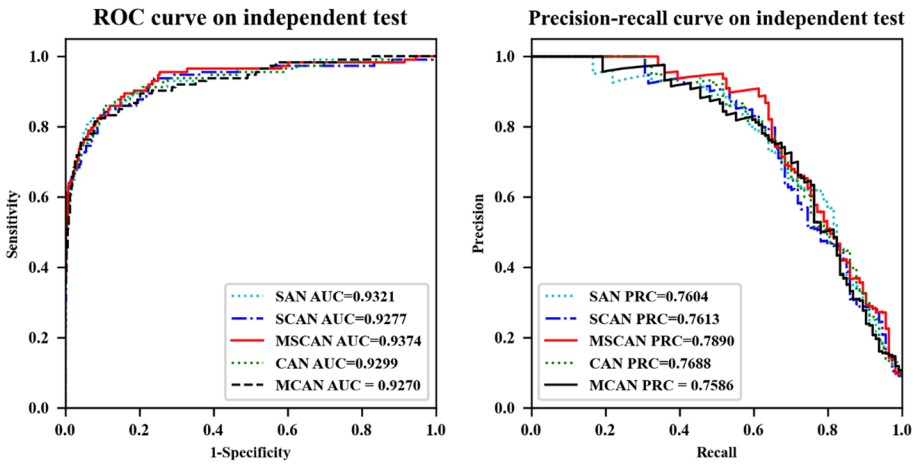


Fig. 2 Performance of MSCAN and variant model on the test data

Table 4 Comparing MSCAN and variant model evaluation results based on test data of Chen et al.

Classifiers	AUROC	ACC	Sen	Precision	MCC	Spe	F-1	AUPRC
SAN	0.9321	95.05	55.26	85.14	66.24	99.04	67.02	0.7604
SCAN	0.9277	95.29	53.51	**91.04**	67.73	**99.47**	67.40	0.7613
MSCAN	**0.9374**	**95.69**	**58.77**	90.54	**70.95**	99.39	**71.28**	**0.7890**
CAN	0.9299	94.89	52.63	85.71	64.81	99.12	65.21	0.7688
MCAN	0.9270	94.81	55.26	81.82	64.71	98.77	65.97	0.7586

Bold Indicates the best performance

SAN contains only self-attention; SCAN, and MSCAN are combinations of self- and cross-attention; CAN and MCAN are combinations of only cross-attention

our findings emphasize the value of the multi-scale self- and cross-attention approach employed by MSCAN in advancing the understanding of RNA modifications and their functional implications.

Comparison with state-of-the-art approaches

We compared MSCAN with several state-of-the-art models, including m6A-word2vec, DeepM6ASeq, and Plant6mA. To ensure robust evaluation, we employed a fivefold cross-validation on the training data of Chen et al. As shown in Fig. 3 and Table 5, Our results demonstrate that MSCAN outperforms the other models, substantially improving prediction accuracy.

In particular, MSCAN achieves a 4.84% enhancement in the AUPRC metric compared to the second-best performing model, Plant6mA. This superior performance can be attributed to utilizing the multi-scale self- and cross-attention mechanisms in MSCAN, as opposed to the self-attention mechanism employed by Plant6mA. The results underscore the effectiveness of MSCAN in identifying RNA methylation sites.

Next, we compare the performance of MSCAN with other state-of-the-art models using the test data of Chen et al. The results, as illustrated in Fig. 4 and summarized in Table 6, demonstrate the superior performance of MSCAN in predicting RNA methylation sites.

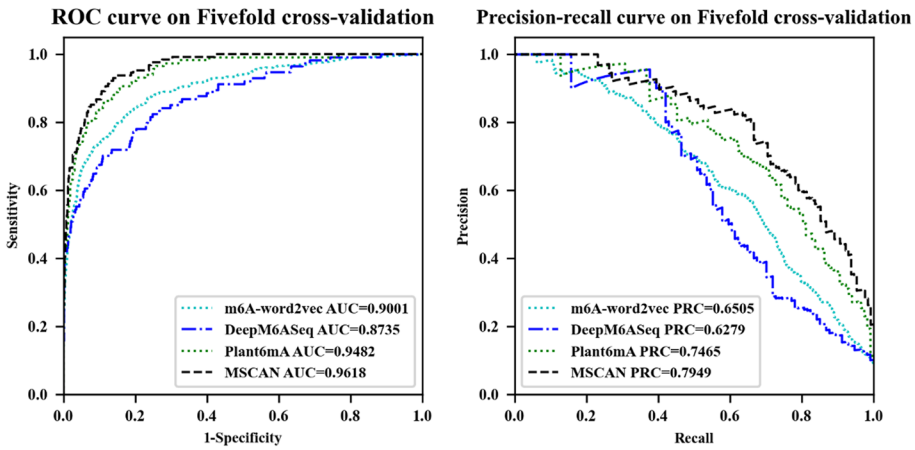


Fig. 3 Performance of the different models on the training data

Table 5 Evaluation results of MSCAN and other state-of-the-art models based on five-fold cross-validation using the training data of Chen et al.

Classifiers	AUROC	ACC	Sen	Precision	MCC	Spe	F-1	AUPRC
m6A-word2vec	0.9001	93.30	53.79	66.18	56.10	97.25	59.35	0.6505
DeepM6ASeq	0.8735	93.38	46.49	70.67	54.02	98.07	56.08	0.6279
Plant6mA	0.9482	93.48	**74.36**	61.27	63.97	95.37	67.18	0.7465
MSCAN	**0.9618**	**94.86**	59.69	**83.70**	**68.11**	**98.72**	**69.68**	**0.7949**

Bold indicates the best performance

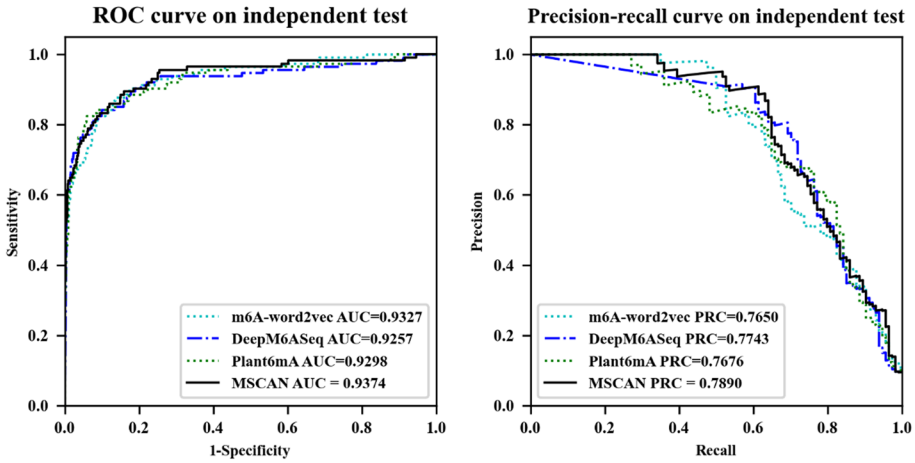


Fig. 4 The ROC and PRC of MSCAN and other state-of-the-art models on the test data

Table 6 Evaluation results of MSCAN and other state-of-the-art models based on the test data of Chen et al.

Classifiers	AUROC	ACC	Sen	Precision	MCC	Spe	F-1	AUPRC
m6A-word2vec	0.9327	93.62	**67.54**	64.17	62.32	90.43	65.81	0.7650
DeepM6ASeq	0.9257	95.37	65.79	79.79	70.00	98.33	**72.12**	0.7743
Plant6mA	0.9298	95.13	56.14	85.33	66.89	99.04	67.72	0.7676
MSCAN	**0.9374**	**95.69**	58.77	**90.54**	**70.95**	**99.39**	71.28	**0.7890**

Bold indicates the best performance

MSCAN outperforms DeepM6ASeq and m6A-word2vec by 1.47% and 2.4% in terms of AUPRC, respectively. This enhanced performance can be attributed to the multi-scale self- and cross-attention network's ability to capture meaningful sequence encodings for more accurate classification. Furthermore, MSCAN surpasses Plant6mA by 2.14% in AUPRC, which may further verify the limitations of the single-scale self-attention mechanism in learning complex contextual relationships between sequence elements. The integration of the cross-attention mechanism enables the model to discern deeper sequence meanings, thus improving its performance.

Assessing model reliability

To evaluate the reliability of our proposed model, we performed one hundred replications of experiments using the test data from Chen et al., evaluating the m6A-word2vec, DeepM6ASeq, Plant6mA, and MSCAN models. In each replication, we used the same test data and ran each model under identical conditions to ensure experimental consistency.

To evaluate the statistical significance of AUPRC values between different methods, we employed Student's t-test [31]. This statistical method helps determine whether performance differences between different methods are significant. Table 7 below shows the *p* values for the difference in the performance of the four classifiers.

Table 7 A statistically significant correlation matrix for the difference in the performance of the four classifiers

Classifiers	Classifiers			
	m6A-word2vec	DeepM6ASeq	Plant6mA	MSCAN
m6A-word2vec				
DeepM6ASeq	0.001243			
Plant6mA	2.905E−44	8.01217E−35		
MSCAN	4.71415E−64	1.25245E−58	0	

Assessing model generalization ability

Based on the data set of Song et al., the generalization ability of MSCAN was evaluated by training the model individually for each methylation type. As presented in Table 8, the MSCAN model consistently outperforms state-of-the-art models, including m6A-word2vec, DeepM6ASeq, and Plant6mA. This result provides empirical evidence of the model's generalizability across diverse methylation site prediction tasks.

Theoretically, the self- and cross-attention mechanism employed by the MSCAN model enables it to capture long-range dependencies and complex interactions between input features more effectively than other models, such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). This characteristic is particularly advantageous in discerning biologically relevant patterns in methylation site prediction, which may contribute to the model's enhanced generalizability.

Comparison with cross-modification validation approaches

Thus far, our results have demonstrated the model's robust classification performance. Notably, a significant advantage of the proposed MSCAN model is its ability to learn the underlying associations among different RNA modifications. Previous studies have revealed clear evolutionary and functional cross-talk among various post-translational modifications of proteins [32] and histone and chromatin modifications [33]. Such associations might also exist at the epi-transcriptome level among different RNA modifications.

To better understand the inherent shared structures among different RNA modifications, we performed cross-modification validation on the second dataset. The resulting AUROC values are displayed in Fig. 5. As the figure shows, cross-modification validation yielded poorer prediction results than those obtained using modification-consistent data and models, indicating the specificity of our method for a particular modification.

Interestingly, in experiments where the test dataset and model were inconsistent, some groups achieved high AUROC values greater than 0.85, suggesting strong and significant positive associations among certain RNA modifications, even those originating from different nucleotides. This observation implies the existence of regions intensively modified by multiple RNA modifications, which likely serve as key regulatory components for the epi-transcriptome layer of gene regulation. Notably, the sequence signatures of these key regulatory regions are largely shared among different RNA modifications (including those that modify different nucleotides) and were successfully captured by our model.

Table 8 Compare MSCAN to other methods under AUC

Classifiers	m^6A	Ψ	m^1A	m^6Am	Am	Cm	Gm	Um	m^5C	m^7G	m^5U	I
m6A-word2vec	0.9773	0.7060	0.8385	0.9867	0.9174	0.9120	0.9554	0.8467	0.9611	0.7505	0.9499	0.6180
DeepM6ASeq	0.9752	0.7510	0.8289	0.9837	0.9213	0.9173	0.9538	0.8716	0.9675	0.7527	0.9584	0.5872
Plant6mA	0.5964	0.5478	0.7268	0.7826	0.8110	0.8016	0.8279	0.7847	0.6806	0.6690	0.9528	0.5137
MSCAN	**0.9834**	**0.7622**	**0.8541**	**0.9904**	**0.9292**	**0.9203**	**0.9577**	**0.8966**	**0.9729**	**0.7994**	**0.9674**	**0.6569**

Bold indicates the best performance

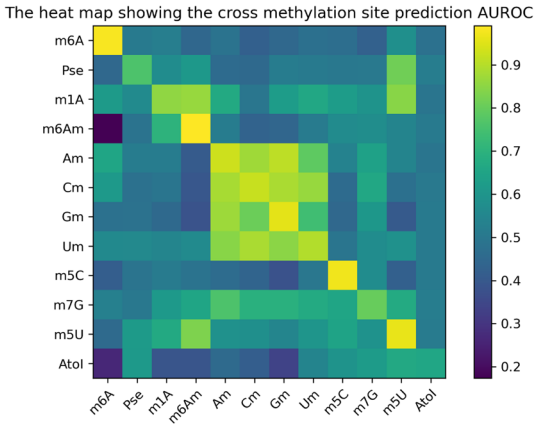


Fig. 5 Heat map of different AUROC values in cross-methylation validation. The horizontal axis is the model type, and the vertical axis is the test data type

Table 9 Association of RNA modifications revealed by MSCAN

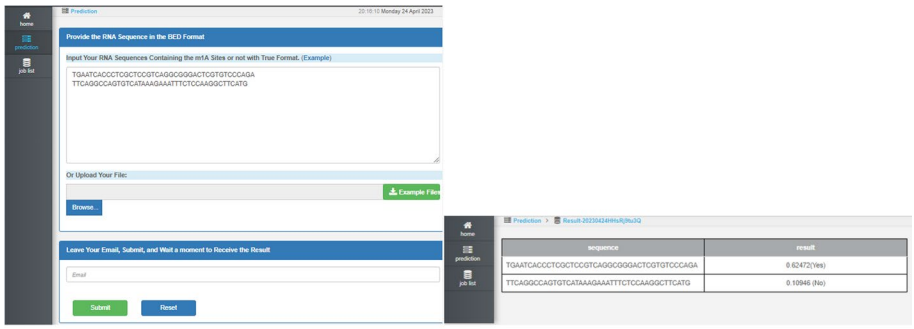
Dataset	Model	AUROC	AUPRC	ACC	Dataset	Model	AUROC	AUPRC	ACC
Am	Am	0.9292	0.9231	86.36	Gm	Gm	0.9577	0.9637	88.33
Gm	Am	0.9082	0.9233	81.11	Am	Gm	0.8695	0.8809	80.16
Cm	Am	0.8724	0.8906	73.51	Cm	Gm	0.8083	0.8280	72.18
Um	Am	0.7884	0.7839	60.00	Um	Gm	0.7387	0.7659	66.34
Cm	Cm	0.9203	0.9273	83.44	Um	Um	0.8966	0.8756	82.92
Um	Cm	0.8647	0.8514	79.75	Cm	Um	0.8859	0.8739	81.78
Am	Cm	0.8830	0.8862	72.31	Am	Um	0.8458	0.8282	78.51
Gm	Cm	0.8865	0.9070	69.44	Gm	Um	0.8480	0.8589	76.66

Table 10 Compare the average bit-score of various methylated sequences

Query subject	Am	Gm	Cm	Um
Am		9.679614	5.866832	3.30773
Gm			3.357072	1.639136
Cm				8.793488
Um				

As presented in Table 9, the most strongly associated modifications originated from the same type of base, with A and G belonging to purine-like bases, and C and U belonging to pyrimidine bases.

To further verify this finding, we compared Am, Gm, Cm, and Um correlations through local BLAST [34] software. First, the Am, Gm, and Cm comparison libraries are established based on the Am data set, Gm data set, and Cm data set respectively. Secondly, the Am, Gm, Cm, and Um data sets are used to compare the comparison libraries with different methylation in pairs. Then, the BLAST output table is obtained. Finally, compare the average value of the comparison result "bit-score". As shown in Table 10, the average bit-score value of the Gm sequence compared to the Am comparison library is high, indicating that the Am sequence and the Gm



a **b**

Fig. 6 Webserver interface. a. Input interface. b. Prediction result

sequence are highly similar. Similarly, the average bit-score value of the Um sequence compared to the Cm comparison library is high, indicating that the Um sequence and Cm sequence similarity is high, which may validate the idea that the most closely related modifications originate from the same type of bases.

Our model provides experimental verification of the existence of an inherent shared structure between different RNA modifications. These findings underscore the potential of the MSCAN model in advancing our understanding of the complex interplay between various RNA modifications and their functional implications.

Web server

We have developed a user-friendly web server for predicting twelve widely occurring human RNA modification sites (m^6A, m^1A, m^5C, m^5U, m^6Am, m^7G, Ψ, I, Am, Cm, Gm, and Um), accessible at http://47.242.23.141/MSCAN/index.php, to facilitate the use of the MSCAN model for RNA methylation site prediction. Take the step of predicting the m^1A methylation site as an example. First, click the "Prediction" button and select the "m^1A" successively. Next, type or paste the RNA sequence, as shown in Fig. 6a. Third, leave your email address in the input box and click the "submit" button. After a calculation period, the prediction results will be displayed in a table, as shown in Fig. 6b. This intuitive web server offers researchers an efficient and convenient platform for employing the MSCAN model in their investigations of RNA modifications and their functional implications.

Discussion

First, based on the test data of Chen et al., we compared the performance of various features based on the MSCAN model, including One-hot encoding, ENAC, and Word2vec. The results reveal that Word2vec outperforms One-hot and ENAC in predicting AUROC and AUPRC. Specifically, the AUPRC of $MSCAN_{word2vec}$ is 5.96% and 5.68% higher than that of $MSCAN_{One\text{-}hot}$ and $MSCAN_{ENAC}$, respectively. These findings are in line with Zhang et al.'s study [20], which highlights that One-hot focuses on local semantic information while ENAC only considers the sequence's nucleic acid composition and position, neglecting more profound semantic information. Conversely, Word2vec captures the contextual semantic information of the sequence, significantly enhancing the model's predictive capability.

Second, based on the test data of Chen et al., we assessed the impact of various MSCAN components by comparing the performance of different MSCAN variants, such as SCAN, SAN, MCAN, and CAN. Experimental results show that MSCAN reduces AUPRC by 3.04% and 2.77% respectively after deleting a self-attention module or a cross-attention module. This finding is consistent with Sun et al.'s study [35], which posits that the removal of self- or cross-attention modules leads to diminished model performance. When both the Multi-Scale and cross-attention modules are removed, the AUPRC of MSCAN decreases by 2.86%. This result aligns with Chen et al.'s study [36], which emphasizes that cross-attention effectively learns multi-scale transformer features for data recognition.

Third, we compared the performance of m6A-word2vec, DeepM6ASeq, Plant6mA, and MSCAN based on the test data of Chen et al. MSCAN's AUROC and AUPRC outperformed the other three state-of-the-art models. In particular, MSCAN surpassed Plant6mA by 2.14% in terms of AUPRC. This study substantiates that the utilization of multi-scale input and cross-attention allows the model to extract diverse features and provide deep semantics, which Plant6mA cannot achieve through information fusion from multiple scales. This conclusion is supported by Guo et al.'s study [37], which demonstrated that multi-scale transformers could extract rich and robust features from different scale inputs.

Four, To make fair comparisons with m6A-word2vec, DeepM6ASeq, Plant6mA methods, we tested MSCAN on twelve RNA modification datasets(m6A, m1A, m5C, m5U, m6Am, m7G, Ψ, I, Am, Cm, Gm, and Um). The results show that MSCAN outperforms all other competing methods. our predicted results may also be consistent with biological insights, which illustrates that MSCAN has good robustness.

Five, based on the dataset of Song et al., we designed a cross-modification validation experiment in which twelve different methylation models were tested using twelve sets of methylation test datasets, respectively. We discovered that the most strongly associated modifications originated from the same base class, such as A and G belonging to purine-like bases. The AUROC and AUPRC metrics of the Am test set on the Gm prediction model are second only to the Am test set on the similar Am prediction model. This finding is consistent with Song et al.'s study [5], which proposed the existence of an inherent shared structure between different RNA modifications.

Lastly, we compared Am, Gm, Cm, and Um correlations through local BLAST software. We found the average bit-score value of the Gm sequence compared to the Am comparison library is high, indicating that the Am sequence and the Gm sequence are highly similar. Similarly, the average bit-score value of the Um sequence compared to the Cm comparison library is high, indicating that the Um sequence and Cm sequence similarity is high, which may validate the idea that the most closely related modifications originate from the same type of bases.These findings underscore the potential of the MSCAN model in advancing our understanding of the complex interplay between some RNA modifications and their functional implications.

Conclusions

This study presents a novel multi-scale cross-attention network (MSCAN) for predicting RNA methylation sites. By combining multi-scale, self-, and cross-attention mechanisms, MSCAN effectively extracts in-depth features from 41 base pair sequences at various scales. The model outperforms state-of-the-art predictors for all twelve modification sites, demonstrating its strong generalization ability.

Crucially, through the cross-modification validation experiments, our model unveils significant associations among different types of RNA modifications in terms of their related sequence contexts. This finding offers valuable insights into the complex relationships between RNA modifications and their respective sequence environments.

It is worth noting that the data set samples of the MSCAN model have the following conditions: (1) The sample is a 41-nt fixed-length sequence, (2) The methylation site must be in the center of the sequence, (3) The sample sequence must have a label. It may seem that MSCAN may only be tested by this method.We hope that in the future, targeting the characteristics of RNA sequences of different lengths, the model structure is adjusted to better capture and utilize these characteristics, and focusing particularly on studies that investigate the biological functions and regulatory mechanisms of different RNA sequence lengths.

Materials and methods

Datasets

In the present study, the benchmark datasets employed to train and test the proposed methods were gathered from previous works [5, 25]. These datasets encompass twelve distinct types of RNA modifications, namely m^6A, m^1A, m^5C, m^5U, m^6Am, m^7G, Ψ, I, Am, Cm, Gm, and Um from H. sapiens. They can be downloaded from http://47.242.23.141/MSCAN/index.php, and detailed information is provided in Table 11. To maintain consistency, all sequence samples were adjusted to a length of 41-nt, with the modified or unmodified site positioned at the center. In cases where the original sequence length fell short of 41-nt, we employed a padding technique, appending "−" to the head or tail of the sequence, to ensure a uniform length of 41-nt across all samples. The raw RNA datasets are represented as $R_0 = \{x^n\}_{n=1}^N$, where N is the sequence number, and each $x^n \in \mathbb{R}^L$ is an RNA sequence. Each entry $x_i^n \in \{A, C, G, U, '-'\}$ *or* $x_i^n \in \{A, C, G, U\}, i = 1, 2, 3, \ldots, L$, where L is the fixed sequence length. The model training and experimental parameter optimization of MSCAN are based on the dataset of Chen et al., and the evaluation of MSCAN generalization capability is based on the dataset of Song et al. The ratio of positive-to-negative samples of Chen's and Song's datasets was 1:10 and 1:1, respectively, as shown in Table 11. The corresponding sequences were followed by aligning of the sequences according to sequence-logo representations rendered using the WebLogo program [38, 39], As shown in Fig. 7.

Feature encoding representation

Achieving an effective feature encoding representation of the sequence is crucial for improving the evaluation metrics of a model. This study uses Word2vec to transform the sequence into embedded vector representations. Since its introduction in 2013,

Table 11 A statistic of the training and test datasets

Full name	Dataset	Original base	Number of positive	Number of negative	Source of data
1-Methyladenosine	m^1A_train0	A	593	5930	Chen et al.[25]
	m^1A_test0	A	114	1140	
N6-methyladenosine	m^6A_ train	A	41,307	41,307	Song et al.[5]
	m^6A_test	A	5901	5901	
1-Methyladenosine	m^1A_train	A	7357	7357	
	m^1A_test	A	1051	1051	
5-Methylcytidine	m^5C_train	C	5953	5953	
	m^5C_test	C	850	850	
5-Methyluridine	m^5U_train	U	863	863	
	m^5U_test	U	123	123	
N6,2′-*O*-dimethyl adenosine	m^6Am_train	A	1172	1172	
	m^6Am_test	A	167	167	
7-Methylguanosine	m^7G_train	G	605	605	
	m^7G_test	G	86	86	
Pseudouridine	Pse_train	U	1989	1989	
	Pse_test	U	284	284	
2′-*O*-methyladenosine	Am_train	A	848	848	
	Am_test	A	121	121	
2′-*O*-methylcytidine	Cm_train	C	1058	1058	
	Cm_test	C	151	151	
2′-*O*-methylguanosine	Gm_train	G	636	636	
	Gm_test	G	90	90	
2′-*O*-methyluridine	Um_train	U	1438	1438	
	Um_test	U	205	205	
Inosine	I_train	A	5164	5164	
	I_test	A	737	737	

Word2vec has significantly advanced the performance of a wide array of natural language processing (NLP) tasks.

The Word2vec methodology offers two different frameworks for encoding: Skip-gram and Continuous Bag of Words (CBOW). The Skip-gram approach predicts contextual information surrounding a given word, whereas the CBOW model generates an embedding for the target word based on its contextual associations. These embeddings are derived through a neural network application, adeptly capturing the inherent relationships within the data.

We developed an RNA embedding approach by treating RNA sequences as sentences and k consecutive RNA nucleotides (k-mers) as words within these sentences. Mathematically, we define the mapping from single nucleotides to the vector representation of k-mers $f : \sum^L \mapsto Y^{L-k+1}$, which is subsequently fed into the neural network for training. This process results in d-dimensional embedded vectors, denoted by $X_m^n \in \mathbb{R}^{m \times d_m}$, where $m = L - k + 1$, and d_m represents the embedding dimension. Gene2vec [21] demonstrated that 3-mers provide the optimal prediction performance. Consequently, we adopted a 3-mers encoding strategy for the input data. Specifically, we employed a sliding window of size 3-nt to slide 41-nt sample sequences with one stride, generating a sequence of 39 words. Each word corresponds to an index in all possible 3-mer

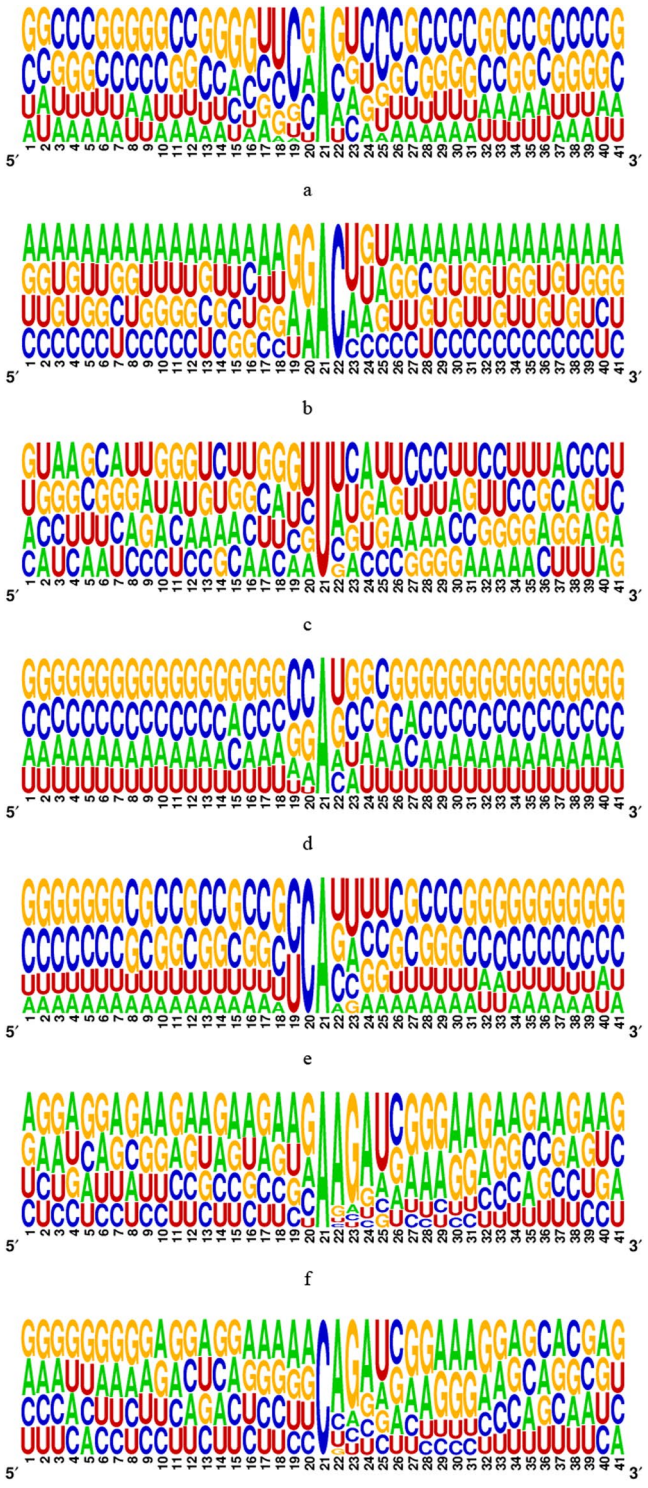


Fig. 7 The motif of methylation sites. **a** m^1A in the dataset of Chen et al. **b** m^6A. **c** Ψ. **d** m^1A. **e** m^6Am. **f** Am. **g** Cm. **h** Gm. **i** Um. **j** m^5C. **k** m^5U. **l** m^7G. **m** I in the dataset of Song et al.

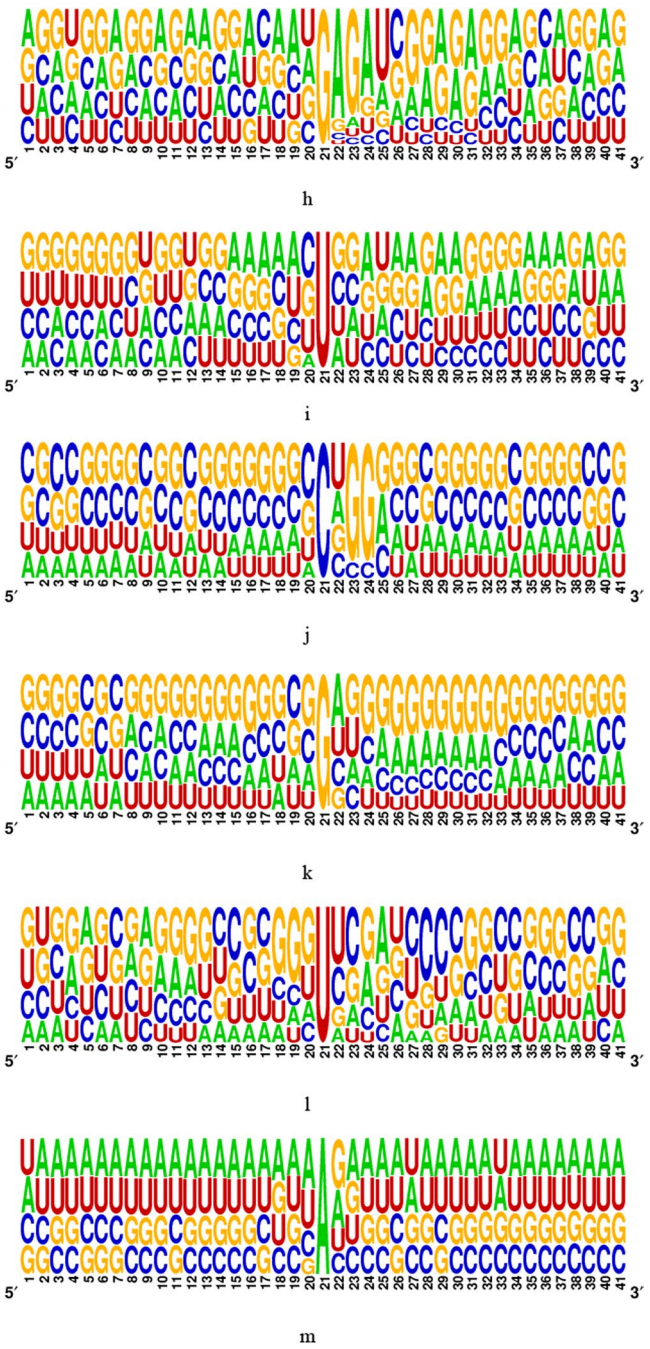


Fig. 7 continued

combinations(105 or 65 types). Given the relatively large dataset and limited word types in our corpus, we chose the Continuous Bag of Words (CBOW) model for encoding, as it offers faster training times than the Skip-gram model. We used the grid-search strategy for the optimization of the parameters for the experiments, word vector dimension in [100,3 00]. Feature encoding with a word vector dimension of 100 achieved the best performance. In summary, each 3-mer is converted into a word vector, transforming a 41-nt sequence into a 39×100 matrix, where 100 represents the word vector dimension.

Model

As shown in Fig. 8, MSCAN represents an innovative DL architecture that employs a combination of multi-scale self- and cross-attention mechanisms and point-wise, fully connected layers in the encoder. This innovative approach enables the effective modeling of both intra- and inter-sequence interactions across a wide range of scales within RNA-seq data by transforming local RNA sequences into high-dimensional vectors via representations through its multi-scale self- and cross-attention networks. MSCAN efficiently extracts crucial RNA sequence features, thereby facilitating the accurate prediction of m^1A modifications.

The results of this study indicate that the nucleotide base neighboring the methylation site is instrumental in determining the specific type of methylation site and its potential functional consequences [40–42]. Therefore, the original sample sequence was extracted with two subsequences. These subsequences were centered on the sequence midpoint. One subsequence was 21-nt long, and the other was 31-nt long, as shown in Fig. 9.

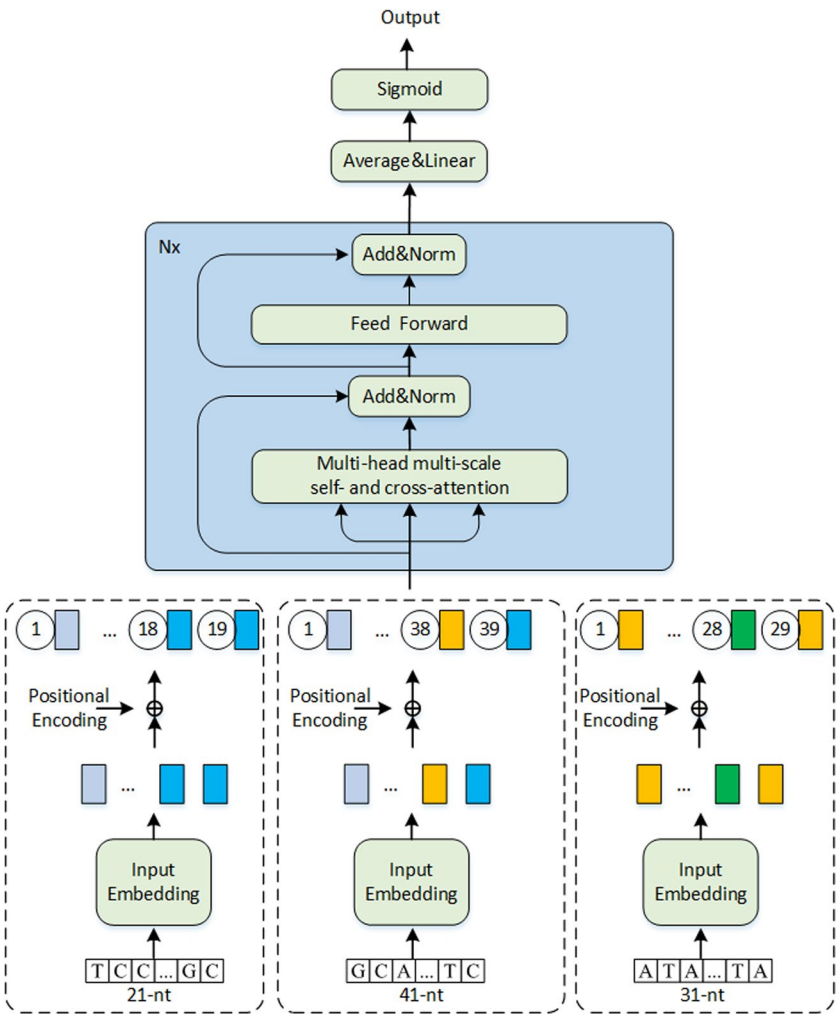


Fig. 8 Structure of our computational framework based on multi-scale self- and cross-attention network to predict m^1A methylation site

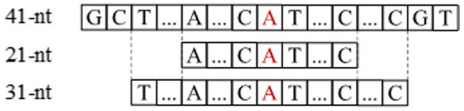


Fig. 9 Schematic diagram of the obtained subsequences

In this paper, we represent the dataset as a collection of sample sequences, each consisting of a main sequence and two subsequences. The dataset can be expressed as $\{(x_{s0}^1, x_{s1}^1, x_{s2}^1, y^1), (x_{s0}^2, x_{s1}^2, x_{s2}^2, y^2), \cdots, (x_{s0}^n, x_{s1}^n, x_{s2}^n, y^n)\}$, where $y^n \in \{0, 1\}$, $x_{s0}^i, x_{s1}^i, x_{s2}^i$ are the three sequences of the i-th sample, x_{s0}^i is the main sequence, with s0 = 41, x_{s1}^i, x_{s2}^i is the subsequence, with s1 = 21, and s2 = 31. Experiments show that the performance of trained models exhibits variability when the order of input sample sequences is altered, as shown in Table 1. MSCAN employs the Word2vec encoder to encode word vectors for these sequences. For example, sequences with lengths 21-nt, 41-nt, and 31-nt are transformed into three distinct matrices of varying dimensions: 19×100, 39×100, and 29×100, respectively.

To account for the lack of recursion or convolution in the model, it is necessary to incorporate information about the relative positions of tokens within sequences so that the model can utilize sequence order effectively. To achieve this, "position encoding" is added to the Word2vec embedding output, forming the input for the encoder. The positional encoding method employed in this work was first introduced by Vaswani et al. [43] in a machine translation task.

The encoder is composed of a stack of N = 3 identical layers. Each layer has two sublayers. The first sub-layer is a multi-scale self- and cross-attention network, while the second is a position-wise, fully connected feed-forward network. To facilitate effective information flow, each of these sub-layers incorporates a residual connection in conjunction with layer normalization.

The output generated by each sub-layer can be expressed as LayerNorm(x + sublayer(x)), where sublayer(x) represents the function associated with the sub-layer in question. Both the embedding layer and all model sub-layers yield outputs with a dimension of $d_{model} = 64$, allowing for seamless residual connections.

Upon completion of the classification process, a linear transformation followed by a sigmoid function is employed to convert the encoder output into predicted probabilities. We used grid-search to choose the hyperparameters on the training data of Chen et al., specifically, epoch in [50, 100], learning_rate in [5e−4, 5e−2], batch in [5, 10, 20, 60], and dropout in [0.2, 0.5]. Final epoch = 100, learning_rate = 5e−4, batch = 10, and dropout = 0.2 is the optimal hyperparameters.

Multi-scale self- and cross-attention network

The multi-scale self and cross-attention network constitutes the initial layer of the encoder, designed to handle linguistic input at various scales. Utilizing word2vec embeddings, matrices at three distinct scales (take X_{s0}, X_{s1}, X_{s2} as an example) are introduced into the self-attention and cross-attention modules for simultaneous computation. Specifically, X_{s0} is incorporated into the self-attention module, while the two combinations (X_{s0} and X_{s1}, X_{s0} and X_{s2}) are integrated into the cross-attention module. Subsequently,

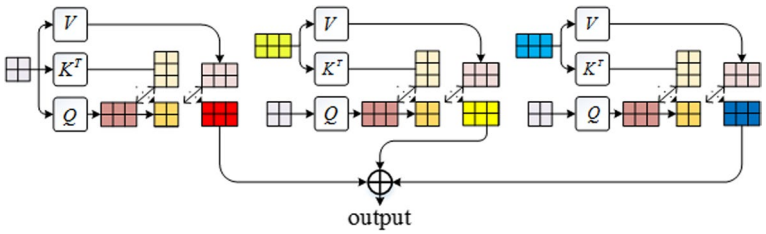


Fig. 10 The internal structure of the multi-scale self- and cross-attention network

the outputs from these modules are directly added and relayed to the subsequent layer, as shown in Fig. 10.

Cross-attention network

The cross-attention network is designed to extract and learn relationships between words in sequences of varying scales, effectively capturing associations across different sequences. Using sequences X_{s0} and X_{s1} as examples, we first transform each sequence into three different terms, which are query, key, and value. This is achieved through the application of linear projections.

$$Q_{s0} = X_{s0}W_{s0}^{Q}, K_{s0} = X_{s0}W_{s0}^{K}, V_{s0} = X_{s0}W_{s0}^{V} \tag{7}$$

$$Q_{s1} = X_{s1}W_{s1}^{Q}, K_{s1} = X_{s1}W_{s1}^{K}, V_{s1} = X_{s1}W_{s1}^{V} \tag{8}$$

where $X_m \in \mathbb{R}^{m \times d_{model}}$ is the output of the sequence embedding module, m represents the length of the input sequence $m \in \{s_0, s_1, s_2\}$. W_m^Q, $W_m^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_m^V \in \mathbb{R}^{d_{model} \times d_v}$. X_m is transformed into the query matrix $Q_m \in \mathbb{R}^{m \times d_k}$, the key matrix $K_m \in \mathbb{R}^{m \times d_k}$, and the value matrix $V_m \in \mathbb{R}^{m \times d_v}$, in which d_k is the dimension of matrices Q_m, K_m, and d_v is the dimension of matrix V_m.

Second, we compute the cross-modal dot product between the query vector of X_{s0} and the key vector of X_{s1}, dividing the result value by $\sqrt{d_k}$, to estimate the association between the X_{s0} and X_{s1}. These results are subsequently refined and normalized utilizing the softmax function, yielding attention weight coefficients. Lastly, we leverage these coefficients to aggregate the corresponding value vectors from each feature sequence, thereby facilitating that the associated information between the two sequences is obtained. The cross-attention function can be described as follows:

$$Cross - Attention(Q_{s0}, K_{s1}, V_{s1}) = softmax\left(\frac{Q_{s0}K_{s1}^{T}}{\sqrt{d_k}}\right)V_{s1} \tag{9}$$

Self-attention network

In contrast to the cross-attention module, which primarily focuses on inter-sequence interactions, the self-attention module identifies and elucidates intra-sequence associations. The self-attention function is described as

$$Self - Attention(Q_{s0}, K_{s0}, V_{s0}) = softmax\left(\frac{Q_{s0}K_{s0}^{T}}{\sqrt{d_k}}\right)V_{s0} \tag{10}$$

Multi-head multi-scale self- and cross-attention

The above elucidation pertains to single-headed attention, a fundamental mechanism in attention-based models. However, multi-headed attention is commonly employed in practice to augment model efficacy and expedite training. This technique entails conducting single-headed attention in parallel across multiple instances, known as "heads", and subsequently integrating the outcomes derived from each head. By incorporating multi-headed attention, the model can effectively capture diverse contextual information and intricate relationships inherent in the input data. The function of cross-attention is described as:

$$\begin{aligned} &MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \\ &\quad where head_i = Attention\left(Q_{s0}W_{s0s0i}^{Q_{s0}}, K_{s0}W_{s0s0i}^{K_{s0}}, V_{s0}W_{s0s0i}^{V_{s0}}\right) \\ &\quad + Attention\left(Q_{s0}W_{s0s1i}^{Q_{s0}}, K_{s1}W_{s0s1i}^{K_{s1}}, V_{s1}W_{s0s1i}^{V_{s1}}\right) \\ &\quad + Attention\left(Q_{s0}W_{s0s2i}^{Q_{s0}}, K_{s2}W_{s0s2i}^{K_{s2}}, V_{s2}W_{s0s2i}^{V_{s2}}\right) \end{aligned} \tag{11}$$

where the $W_{s0s0i}^{Q_{s0}}, W_{s0s0i}^{K_{s0}}, W_{s0s1i}^{Q_{s0}}, W_{s0s1i}^{K_{s1}}, W_{s0s2i}^{Q_{s0}}, W_{s0s2i}^{K_{s2}}, \in \mathbb{R}^{d_{model} \times d_k}$, $W_{s0s0i}^{V_{s0}}, W_{s0s1i}^{V_{s1}}, W_{s0s2i}^{V_{s2}} \in \mathbb{R}^{d_{model} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$

In this task, we employ h = 8 parallel attention layers. For each layer,we use $d_k = d_v = d_{model}/\mathrm{h} = 8$.

Position-wise feed-forward networks

After the multi-headed, multi-scale self- and cross-attention layer, a second sub-layer is incorporated to augment the representative capacity of the model further. This additional component comprises a position-wise, fully connected feed-forward network, enhancing the overall model performance. The architecture of this network entails two successive linear transformations, with an intervening rectified linear unit (ReLU) activation function, ensuring a non-linear and expressive representation of the input data. It is defined as:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{12}$$

The input and output have dimensionality $\mathrm{d_{model}} = 64$, while the inner layer's dimensionality is $\mathrm{d_{ff}} = 256$.

Classification module

To accomplish the classification task, the initial step involves computing the average of the encoder output. Subsequently, a linear transformation is applied, followed by implementing a sigmoid activation function. The optimization of the model is facilitated by employing cross-entropy loss as the primary objective. Finally, the methylation site

probabilities are acquired, providing a robust and comprehensive representation of the underlying biological processes.

Abbreviations

RNA	Ribonucleic acid
m^1A	N1-methyladenosine
m^6A	N6-methyladenosine
Ψ	Pseudouridine
m^6Am	N6,2′-*O*-dimethyl adenosine
Am	2′-*O*-Methyladenosine
Cm	2′-*O*-Methylcytidine
Gm	2′-*O*-Methylguanosine
Um	2′-*O*-Methyluridine
m^5C	5-Methylcytidine
m^7G	7-Methylguanosine
m^5U	5-Methyluridine
I	Inosine
CNN	Convolutional neural network
BiLSTM	Bidirectional long short-term memory
LSTM	Long short-term memory
RNN	Recurrent neural network
NLP	Natural language processing
DL	Deep learning
MSCAN	Multi-scale self- and cross-attention network
MCAN	Multi-scale cross-attention network
SCAN	Self- and cross-attention network
CAN	Cross-attention network
SAN	Self-attention network
Sn	Sensitivity
Sp	Specificity
ACC	Accuracy
Pre	Precision
MCC	Matthews correlation coefficient
AUROC	Area under the receiver operating characteristic
AUPRC	Area under the precision-recall curve
ENAC	Enhanced nucleic acid composition

Acknowledgements
Not applicable.

Author contributions
HW built the architecture for MSCAN, designed and implemented the experiments, analyzed the results, and wrote the paper. LZ and TH conducted the experiments and revised the paper. DW conducted the experiments, analyzed the results, and revised the paper. LZ and YS supervised the project, analyzed the result, and revised the paper. All authors read, critically revised, and approved the final manuscript.

Funding
This work has been supported by the National Natural Science Foundation of China (31871337 and 61971422 to LZ), and the "333 Project" of Jiangsu(BRA2020328 to WHL). The funding body did not play any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
The data supporting the findings of the article is available at the web server http://47.242.23.141/MSCAN/index.php.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 24 April 2023 Accepted: 11 January 2024
Published online: 17 January 2024

References

1. El Allali A, Elhamraoui Z, Daoud R. Machine learning applications in RNA modification sites prediction. Comput Struct Biotechnol J. 2021;19:5510–24.
2. Wang H, Wang SY, Zhang Y, Bi SD, Zhu XL. A brief review of machine learning methods for RNA methylation sites prediction. Methods. 2022;203:399–421.
3. Liu L, Song B, Ma J, Song Y, Meng J. Bioinformatics approaches for deciphering the epitranscriptome: recent progress and emerging topics. Comput Struct Biotechnol J. 2020;18:1587–604.
4. Chen LF, Tan XQ, Wang DY, Zhong FS, Liu XH, Yang TB, Luo XM, Chen KX, Jiang HL, Zheng MY. TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. Bioinformatics. 2020;36(16):4406–14.
5. Song ZT, Huang DY, Song BW, Chen KQ, Song YY, Liu G, Su JL, de Magalhaes JP, Rigden DJ, Meng J. Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications. Nat Commun. 2021;12(1):1–11.
6. Grozhik AV, Olarerin-George AO, Sindelar M, Li X, Jaffrey SR. Antibody cross-reactivity accounts for widespread appearance of m1A in 5′UTRs. Nat Commun. 2019;11:1–13.
7. Dominissini D, et al. The dynamic N(1)-methyladenosine methylome in eukaryotic messenger RNA. Nature. 2016;530(7591):1–39.
8. Chen K, Lu ZK, Wang X, Fu Y, Luo GZ, Liu N, Han DL, Dominissini D, Dai Q, Pan T, et al. High-resolution N-6-methyladenosine (m(6)A) map using photo-crosslinking-assisted m(6)A sequencing. Angew Chem Int Ed. 2015;54(5):1587–90.
9. Li X, Xiong X, Wang K, Wang L, Shu X, Ma S, Yi C. Transcriptome-wide mapping reveals reversible and dynamic N(1)-methyladenosine methylome. Nat Chem Biol. 2016;12(5):311–6.
10. Masiello I, Biggiogera M. Ultrastructural localization of 5-methylcytosine on DNA and RNA. Cell Mol Life Sci. 2017;74(16):3057–64.
11. Xiaoyu L, Xushen X, Meiling Z, Kun W, Ying C. Base-resolution mapping reveals distinct m1A methylome in nuclear- and mitochondrial-encoded transcripts. Mol Cell. 2017;68(5):993–1005.
12. Zhou H, Rauch S, Dai Q, Cui X, Dickinson BC. Evolution of a reverse transcriptase to map N1-methyladenosine in human messenger RNA. Nat Methods. 2019;16(12):1–8.
13. Zhou Y, Zeng P, Li Y-H, Zhang Z, Cui Q. SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. Nucleic Acids Res. 2016;44(10):e91–e91.
14. Chen W, Feng P, Tang H, Ding H, Lin H. RAMPred: identifying the N(1)-methyladenosine sites in eukaryotic transcriptomes. Sci Rep. 2016;6:1–8.
15. Chen W, Feng P, Yang H, Ding H, Lin H, Chou KC. iRNA-3typeA: identifying three types of modification at RNA's adenosine sites. Mol Ther Nucleic Acids. 2018;11:468–74.
16. Liu K, Chen W. iMRM: a platform for simultaneously identifying multiple kinds of RNA modifications. Bioinformatics. 2020;36(11):3336–42.
17. Iuchi H, Matsutani T, Yamada K, Iwano N, Sumi S, Hosoda S, Zhao ST, Fukunaga T, Hamada M. Representation learning applications in biological sequence analysis. Comput Struct Biotechnol J. 2021;19:3198–208.
18. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Mol Syst Biol. 2016;12(7):1–16.
19. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019;51(1):12–8.
20. Zhang L, Li GS, Li XY, Wang HL, Chen ST, Liu H. EDLm(6)APred: ensemble deep learning approach for mRNA m(6)A site prediction. BMC Bioinform. 2021;22(1):1–15.
21. Zou Q, Xing PW, Wei LY, Liu B. Gene2vec: gene subsequence embedding for prediction of mammalian N-6-methyladenosine sites from mRNA. RNA. 2019;25(2):205–18.
22. Xiang S, Yan Z, Liu K, Zhang Y, Sun Z. AthMethPre: a web server for the prediction and query of mRNA m(6)A sites in *Arabidopsis thaliana*. Mol Biosyst. 2016;12(11):3333–7.
23. Lv ZB, Ding H, Wang L, Zou Q. A convolutional neural network using dinucleotide one-hot encoder for identifying DNA N6-methyladenine sites in the rice genome. Neurocomputing. 2021;422:214–21.
24. Tahir M, Hayat M, Chong KT. Prediction of N6-methyladenosine sites using convolution neural network model based on distributed feature representations. Neural Netw. 2020;129:385–91.
25. Chen Z, Zhao P, Li F, Wang Y, Smith AI, Webb GI, Akutsu T, Baggag A, Bensmail H, Song J. Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. Brief Bioinform. 2019;21(5):1676–96.
26. Huang Y, He NN, Chen Y, Chen Z, Li L. BERMP: a cross-species classifier for predicting m(6)A sites by integrating a deep learning algorithm and a random forest approach. Int J Biol Sci. 2018;14(12):1669–77.
27. Zhang Y, Hamada M. DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning. BMC Bioinform. 2018;19:1–11.
28. Tao S, Xz A, Mao DB, Rp C, Sw A, Gan WA. DeepFusion: a deep learning based multi-scale feature fusion method for predicting drug–target interactions. Methods. 2022;204:269–77.
29. Kim Y, Denton C, Hoang L, Rush AM. Structured attention networks. 2017, p. 1–21.
30. Shi H, Li S, Su X. Plant6mA: a predictor for predicting N6-methyladenine sites with lightweight structure in plant genomes. Methods (San Diego, Calif). 2022;204:1–6.
31. Chen Z, Zhao P, Li C, Li FY, Xiang DX, Chen YZ, Akutsu T, Daly RJ, Webb GI, Zhao QZ, et al. iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. Nucleic Acids Res. 2021;49(10):e60.
32. Huang KY, Lee TY, Kao HJ, Ma CT, Lee CC, Lin TH, Chang WC, Huang HD. dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications. Nucleic Acids Res. 2019;47(D1):D298–308.
33. Lee JS, Smith E, Shilatifard A. The language of histone crosstalk. Cell. 2010;142(5):682–5.

34. Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, Madden TL, Matten WT, McGinnis SD, Merezhuk Y, et al. BLAST: a more efficient report with usability improvements. Nucleic Acids Res. 2013;41(W1):W29–33.
35. Sun LC, Liu B, Tao JH, Lian Z. IEEE: multimodal cross- and self-attention network for speech emotion recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP): Jun 06–11 2021; Electr Network. 2021, p. 4275–4279.
36. Chen CF, Fan Q, Panda R. CrossViT: cross-attention multi-scale vision transformer for image classification. In: ICCV. 2021, p. 1–12.
37. Guo Q, Qiu X, Liu P, Xue X, Zhang Z. Multi-scale self-attention for text classification. In: Proceedings of the AAAI conference on artificial intelligence, 2020, p. 7847–7854.
38. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188–90.
39. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990;18(20):6097–100.
40. Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD, Yu M, Ecker JR. Global epigenomic reconfiguration during mammalian brain development. Science. 2013;341(6146):629.
41. Guo JU, Su Y, Shin JH, Shin J, Li H, Xie B, Zhong C, Hu S, Le T, Fan G. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. Nat Neurosci. 2014;17(2):215–22.
42. Ziller MJ, Müller F, Liao J, Zhang Y, Gu H, Bock C, Boyle P, Epstein CB, Bernstein BE, Lengauer T. Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. PLoS Genet. 2011;7(12):e1002389.
43. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. arXiv. 2017, p. 1–15.

Publisher's Note