GLF6mA: A novel model for N6-methyladenine DNA methylation identification with gene expression data

Qing Wang[1, 2, †], Lin Song[3, †], Weiping Liu[1, 2], Xinghong Chen[1, 2], Xiumei Wang[1, 2], Bin Yang[1, 2], Juhao Jiang[1, 2], Xuran Zhou[1, 2], Guannan Chen[1, 2, *] and Yao Lin[4, *]

[1] Key Laboratory of OptoElectronic Science and Technology for Medicine of Ministry of Education, Fujian Provincial Key Laboratory of Photonics Technology, Fujian Normal University, Fuzhou 350007, China
[2] Fujian Provincial Engineering Technology Research Center of Photoelectric Sensing Application, Fujian Normal University, Fuzhou 350007, China
[3] Department of Epidemiology and Medical Statistics, School of Public Health, Nantong University, Nantong 226001, China
[4] Central Laboratory at the Second Affiliated Hospital of Fujian Traditional Chinese Medical University, Innovation and Transformation Center, Fujian University of Traditional Chinese Medicine, Fuzhou 350122, China

[*]Corresponding authors: **Guannan Chen**, Email: edado@fjnu.edu.cn; **Yao Lin**, Email: yaolin@fjtcm.edu.cn.
[†]These authors contributed equally to this work.

Abstract

N6-methyladenine (6mA) refers to the methylation modification of the sixth nitrogen atom on adenine. It plays an essential role in maintaining the normal transcriptional activity of cells, DNA damage repair, chromatin remodeling, genetic imprinting, embryonic development and tumorigenesis. Traditional experimental techniques are time-consuming, costly and labor-intensive, so it is difficult to identify 6mA sites from high-throughput sequences. Some machine learning models based on computational methods can handle the identification of 6mA sites in multiple sequences at the same time, which is time-saving, efficient and labor-saving. As an effective supplement to biological experimental method, it is gradually favored by the biological community. However, the existing methods of 6mA sites identification are often limited in following issues, including employed single feature and showed poor performance of cross species identification in several datasets of gene expression. To address these issues, in this paper, we proposed a novel 6mA sites identification model, termed GLF6mA, which is composed of graph convolutional network (GCN), long short term memory (LSTM) and fully convolutional network (FCN). We evaluated GLF6mA on four public datasets (Rice, F. vesca, E. coli and A. thaliana), GLF6mA demonstrated high prediction performance, and the accuracy is 92.61%, 98.26%, 93.42% and

96.93% in four public datasets, respectively. Our results showed that GLF6mA model has strong 6mA site prediction ability and thus providing a new insight and tool for disease diagnosis and clinic practice.

Key words: N6-methyladenine (6mA), DNA Methylation, Gene Expression, Graph Neural Network, Disease Diagnosis, Clinical

1. Introduction

DNA methylation refers to the process in which DNA molecules are connected to methyl groups with S-AdenosylMethionine (SAM) as the methyl donor, catalyzed by DNA methyltransferase [1-3]. DNA methylation is an important genetic modification that can be inherited, and it plays an important role in biological evolution [4, 5]. Due to the limited level of experimental technology, DNA methylation, which was widely studied in the past, refers to the methylation of the fifth carbon atom of cytosine to form 5-methylcytosine (5mC) [6-8]. In recent years, with the discovery of the methylation modification of the sixth nitrogen atom of adenine, N6-methyladenine (6mA) [9], has overturned the previous research belief of community that only 5mC is methylated in mammals. In some views, 6mA was first discovered in Escherichia coli [9, 10]. With the improvement of technology, 6mA has been detected in rice, corn and human cells. As an important non-permanent but relatively long-term heritable genetic modification, 6mA methylation has been found to maintain normal cell transcriptional activity, DNA damage repair ability, chromatin remodeling, genetic imprinting, embryonic development, and tumors become a research hotspot in the field of molecular biology and medicine [11, 12].

Accurate identification of methylation sites in sequences is of great significance for the development of new drugs and the regulation of human biochemical characteristics. With the implementation of the Human Genome Project, the sequencing technology has been developed rapidly. Researchers have successfully obtained drafts of biological genomes such as rice, peanut, corn, wheat, fruit flies, fungi, bacteria, and mice [13]. However, the expression richness of 6mA at the DNA sequence is relatively low. In mammals, there are less than 10 6mA sites per million adenines. The identification of 6mA sites is complicated, and it is not feasible to detect 6mA sites from large amount of sequences. Till now only few 6mA map of genomes have been identified. There are already several methods to identify 6mA sites, such as methylated DNA immunoprecipitation sequencing, capillary electrophoresis and laser induced fluorescence (CE-LIF) and single molecule real-time

sequencing (SMRT-seq) [14, 15]. Recently, through mass spectrometry analysis and 6mA immunoprecipitation followed by sequencing, the 6mA map of the rice genome has been identified. Although it is possible to identify methylation site through experimental methods, due to the low amount of methylation, it can only be detected by advanced technology, which is time-consuming, labor-intensive and very expensive, and relies solely on experimental methods [16, 17]. The computer method of methylation identification can identify and predict 6mA sites from large-scale sequences with high identification efficiency, which has great practical significance for reducing costs of experiments and for in-depth experimental research on methylation [18].

In recent years, the rise and application of neural networks have promoted the research of pattern recognition and data mining [19]. Many machine learning methods that once relied on manual feature extraction (such as target detection, machine translation, and speech recognition) have now been solved by various end-to-end deep learning models. Although traditional deep learning methods have been applied to extract the features of Euclidean spatial data effectively, the data in many practical application scenarios are generated from non-Euclidean spaces [20, 21]. The performance of traditional deep learning methods is not very successful in processing non-Euclidean spatial data. Recently, the deep learning methods based on graphs for non-Euclidean spatial data analysis have developed rapidly. Researchers proposed convolutional neural networks (CNN), recurrent neural networks (RNN), long short term memory (LSTM) model and deep autoencoders to define and design graph neural networks (GNN) for processing graph data [22, 23]. In order to improve the efficiency for identifying and predicting 6mA sites across species, through the analysis and evaluation of the existing DNA sequence characteristics and single machine learning methods, an integrated learning method (GLF6mA) with better performance was constructed based on the rice dataset; GLF6mA displayed better in F1-Score (F1), Matthews correlation coefficient (MCC), Sensitivity (Sn), Specificity (Sp), Accuracy (ACC), Area under ROC curve (AUC) and Area under PR curve (AUPRC) using several public DNA 6mA datasets, including F. vesca, Rice, E. coli and A. thaliana. GLF6mA effectively solves the problem of 6mA sites identification and prediction from large-scale sequences and will promote the research of 6mA sites in molecular biology and medicine, and it is conducive to the diagnosis of cell development or some disease prevention and treatment.

The main contributions of the paper are summarized as follows:

(1) We explored how to establish a prediction model for DNA methylation sites. At present, gene expression data is the focus of main research on cancer diagnosis and classification

using machine learning algorithms. Gene expression is closely related to the occurrence and development of cancer. However, as an important way of epigenetic modification, DNA methylation plays a vital role in the biological activities of cells, and the status of abnormal methylation is also related to cancer. The formation and development are closely related. Therefore, we adopted deep learning methods to establish a DNA 6mA sites identification model based on DNA methylation data, thus contributing new ideas and methods to disease research and diagnosis.

(2) We proposed a novel 6mA site prediction method on the basis of multi-model fusion and GNN to classify the data of gene expression. Since each deep learning algorithm has the advantages and disadvantages of their own, multi-model fusion can be performed to complement the advantages of the model while improving its prediction performance. The use of one-dimensional deep CNN to predict gene expression data is better at fully mining the inherent characteristics of the original data and improving the prediction performance. In our study, the prediction performance of the proposed model was evaluated on four public datasets, namely, E. coli, F. vesca, A. thaliana and Rice.

(3) There were plenty of experiments conducted on these four public datasets. At the same time, various common evaluation metrics (Sn, Sp, AUC, AUPRC, F1, MCC and ACC) were applied to verify the performance of the model. According to the experimental results, our proposed model outperforms other state-of-the-art methods in making prediction on both four public datasets. GLF6mA can be used not only for the identification of DNA 6mA sites (prediction), but also for the prediction of other DNA methylation sites, such as 5mC.

All the data is available and can be downloaded at https://drive.google.com/file/d/1VMl3rNU9UhCWmP4ISdMOPCOqtJ_1RD3T/view?usp=sharing.

2. Materials and Methods

2.1 Dataset

In this study, a benchmark dataset comprised of 6mA sites of four species was used, the information of which is shown in Table 1. It includes the 6mA dataset in the rice DNA sequence that has been used. Obtained from the National Center for Biotechnology Information (NCBI) [24], the dataset contains 880 experimentally verified sequence fragments of 6mA sites and 880 non-6mA site sequence fragments. The sequence length is 41nt, the F. vesca dataset was collected from the MDR database [25], including 4303 6mA site sequence fragments and 4303 non-6mA site sequence fragments, and the DNA sequence

length is also 41nt, as obtained from UCI database [26]. The E. coli 6mA site dataset contains 388 6mA site sequence fragments and 388 non-6mA site sequence fragments, while the sequence length is also 41nt [27]. Finally, the 6mA site data of A. thaliana was collected from the MethSMRT database [28]. The dataset contains 2100 6mA site sequence fragments and 2100 non-6mA site sequence fragments, while the sequence length is 101nt. The details of both datasets are shown in Table 1 below.

Table 1 The detail information of some DNA 6mA public datasets.

Dataset	Positive	Negative	Length	Total
Rice	880	880	41nt	1760
F. vesca	4303	4303	41nt	8606
E. coli	388	388	41nt	776
A. thaliana	2100	2100	101nt	4200

2.2 Long Short Term Memory (LSTM)

Long short term memory (LSTM) network represents an extraordinary RNN model. It performs has now been widely applied for the excellent performance in addressing various sequence problems. Its exceptional structure design enables it to solve the long dependence task, which is to conductive to extract the long pattern information hidden in the sequence sample data. As the critical part of LSTM, the cell state is relied on to save the current state information of the LSTM for transfer to the LSTM next time. In LSTM, the specially designed "gates" are used either to introduce or remove the information in cell state c_t. A gate provides a means for information to pass it selectively. For LSTM, there are three main gate structures: forgetting gate, memory gate and output gate. These three gates are used to control the retention and transmission of information for LSTM. Ultimately, the cell state c_t and output state h_t are reflected. The basic structure of LSTM is illustrated in Fig. 1.

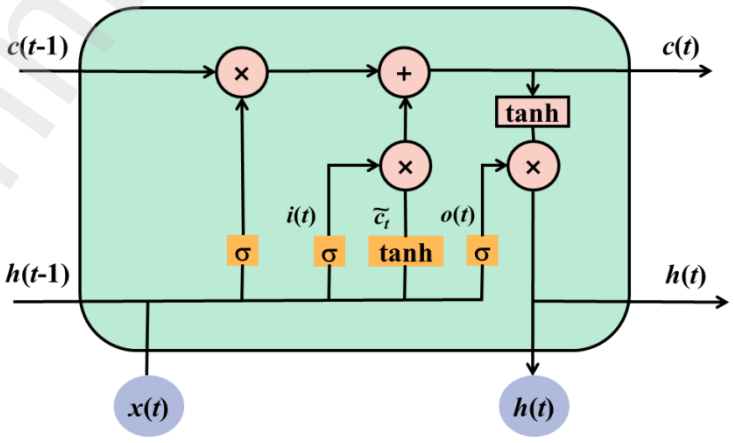


Fig. 1. The basic structure of LSTM, where σ is sigmoid activation function, tanh is hyperbolic tangent function, c_t is the current state information of memory unit, h_t is the current state output, respectively.

2.3 Graph Neural Network (GNN)

Currently, convolutional neural network (CNN) has been extensively applied in various settings like image classification, semantic segmentation, machine translation and many more. Although the underlying data representation in these fields is all grid-like [29], it is impossible for the data involved in various interesting tasks to be represented by a grid-like structure. Instead, it is located in an irregular domain, for example, three-dimensional grids, social networks, telecommunications networks, biological networks, and brain connections [30]. In general, these data can be represented in the form of graphs.

As a kind generalization of RNN, GNN is capable to process more general graph data directly, such as recurrent graphs, directed graphs, and undirected graphs. The graph neural network involves an iterative process. In the first iteration, the process propagates the state of the nodes to the equilibrium state. Then, a neural network is used to generate the final output depending on the state of each node. In recent years, GNN has been widely applied in various fields, such as social networks, knowledge graphs, recommendation systems, and even life science [31]. Since GNN performs well in modeling the dependencies between graph nodes, this has contributed to a major breakthrough in the research of graph analysis [32]. However, there are some downsides to the original GNN:

(1) If the assumption of "fixed point" is dropped, then a more stable representation can be learned using multilayer perceptrons and the need for iterative update process can be eliminated. This is because conversion functions are involved in different iterations. Although the parameters of f are the identical, the parameters in different layers of MLP vary, which allows hierarchical feature to be extracted;

(2) It is incapable to process edge information. For example, the different edges in the knowledge graph may represent different relationships between nodes;

(3) Fixed points may have adverse effect on the diversity of node distribution and they are unfit for representing nodes.

2.4 Graph Convolutional Network (GCN)

Over the past few years, CNN has been experienced rapid advancement, with wide spread attention drawn to their excellent modeling capabilities. Compared with those traditional methods, CNN has achieved a satisfactory performance in image processing and natural language processing (NLP), such as machine translation, image recognition and speech recognition [33]. Besides, traditional convolutional neural networks are restricted to

processing only European spatial data (such as images, text, and speech, etc.), which leads to translation invariance.

In the meantime, a sort of non-European data, graph data, has increasingly attracted attention because of its ubiquity. In practice, graph data structure can represent some data, such as transportation networks, the world wide web (WWW) and social networks [34]. Different from image or text, each node in the graph data has a different local structure to others, which makes the translation invariance no longer satisfied. Due to the lack of translation invariance, it is difficult to define CNN on graph data. In recent years, the ubiquity of graph data has made, researchers pay attention to exploring how deep learning models can be constructed using graph data. As a variety of neural network, GCN has been receiving popularity in recent years [35]. In comparison with the traditional network models, such as LSTM and CNN, which can be used solely for grid structure data, GCN is more capable to process data with non-European structure, such as social network, protein structure and molecular structure, etc.

In this paper, the spectral domain graph convolution method is adopted for the graph network. In the convolution of spectral domain graph, the graph is represented by the corresponding Laplacian matrix $\boldsymbol{L}$:

$$L = D^{-\frac{1}{2}}(D-A)D^{-\frac{1}{2}} = I_N - D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \tag{1}$$

$$D_{ij} = \sum_i A_i \tag{2}$$

In Equation (2), $\boldsymbol{D}$ represents the degree matrix. $\boldsymbol{I_N}$ indicates the identity matrix of size $N \times N$. By decomposing $\boldsymbol{L}$ into eigenvalue, it can be obtained that:

$$L = U\Lambda U^T \tag{3}$$

In Equation (3), $\boldsymbol{\Lambda}$ refers to a diagonal matrix comprised of the eigenvalues of $\boldsymbol{L}$. $\boldsymbol{U} = \{u_1, u_2, ..., u_N\}$, which is comprised of the eigenvectors of $\boldsymbol{L}$, corresponding to a set of orthogonal bases in the $\boldsymbol{R}^N$ space. Graph convolution relies on the diagonalized linear operators defined in the Fourier domain to equivalently replace the classic convolution operators for convolution operations. The convolution kernel g_θ is applied to convolve the graph G:

$$g_\theta * x = Ug_\theta U^T x \tag{4}$$

Equation (4) is suitable for the calculation of small-scale graph structure. In case of a large graph structure, it can be solved by the approximate expansion of Chebyshev polynomials. In this paper, the inter-layer propagation formula using spectral domain convolution is expressed as:

$$H^{l+1} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^l W^l) \quad (5)$$

$$\tilde{A} = A + I_N \quad (6)$$

$$\tilde{D} = \sum_i \tilde{A}_{ij} \quad (7)$$

$$H^l \in R^{N \times \cdot} \quad (8)$$

$$W^l \in R^{F \times \cdot} \quad (9)$$

In Equation (5), σ represents the activation function. As for the self-loop operation on the adjacency matrix by adding the identity matrix, the characteristics of the central node itself can be involved in the convolution; H^l indicates the activation value of the l-th layer; $H^0 = X$, where X denotes the initial input value. W^l represents the learning parameters of the l-th layer.

Given our aim of 6mA sites identification, the bases are defined as nodes, while the connections between bases are defined as edges, with genes abstracted into an undirected graph. The graph is defined as G (V, E, $\boldsymbol{A}$), where V refers to the vertex. That is to say, the base in the gene, E is the edge, which means the association between bases. $\boldsymbol{A}$ is the adjacency matrix, which shows the adjacency between base pairs.

For the construction of the adjacency matrix, most researchers prefer to determine the element value corresponding to the adjacency matrix based on the Euclidean distance or Mahalanobis distance between nodes. According to this method, the constructed adjacency matrix is invariant, which imposes limit on the model parameter update. Therefore, we adopt the gradient update for the model to update the adjacency dynamically matrix according to the loss function during the training process, which makes the graph structure correctly enough.

2.5 Our Method (GLF6mA)

As shown in Fig. 2, the GLF6mA model consists of a GCN model and a LSTM unit. With the sequence data of lengths factored into the model, the double-layer GCN is used to analyze the topology of the grid monitoring data, extract spatial features, and convert the original data into time series data with spatial features. On this basis, the time series data with spatial features are inputted into the LSTM for the extraction of time features. Ultimately, the prediction results are obtained through the fully connected network (FCN) and the log_softmax layer.

Firstly, the raw data are preprocessed and divided. Then, in order to train the model, the

training data consisting of the training set and the verification set are inputted into the GLF6mA model as shown in Fig. 2. Ultimately, the test data are inputted into the trained model, for establishing whether the prediction accuracy of the model meets the expectation standard. If it falls short of the expectation standard, the number of hidden layer nodes in the model would be adjusted and these nodes would be trained again until the model achieves the accuracy is expected.

The loss function for training our proposed model is treated as the binary cross-entropy function to measure the difference between the prediction and the target.

$$L(w) = -\sum_{i=1}^{N} y_i \log(y_i^{'}) + (1 - y_i) \log(1 - y_i^{'}) + \alpha \left\| w \right\|_2 \tag{10}$$

where y_i represents the true label, $y_i^{'}$ indicates the predicted value of each label, and α is intended to avoid over-fitting through perform regularization.

Based on comprehensive temporal and spatial characteristics, the GLF6mA model is capable to produce more accurate prediction results for future air quality, which is advantageous over other models.

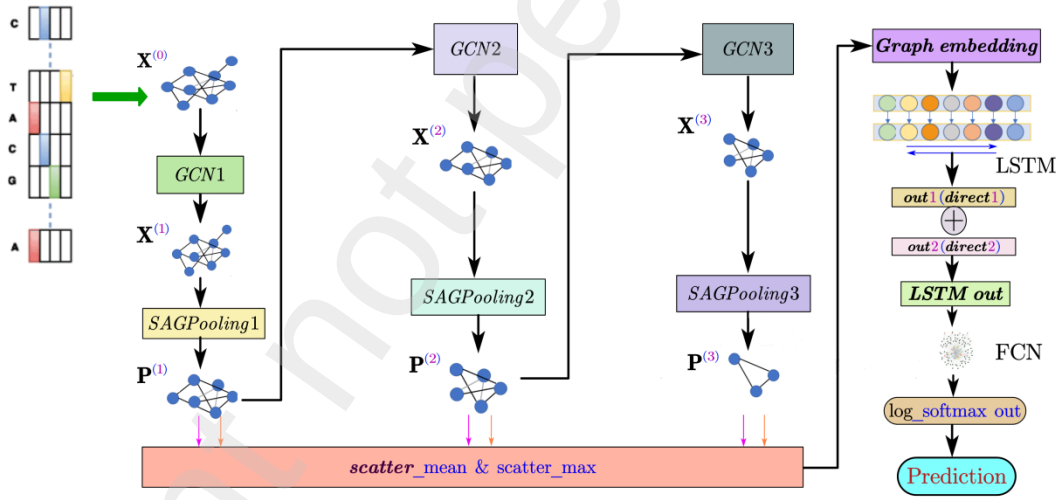


Fig. 2. The flow chart of GLF6mA. The structure of GLF6mA consists of GCN, LSTM (two layers and two directions) and Fully Convolutional Network (FCN), where $\boldsymbol{X}^{(i)}$ represents $\{X_1^{(i)}, X_2^{(i)}, \ldots, X_k^{(i)}\}$ and $\boldsymbol{P}^{(i)}$ represents $\{P_1^{(i)}, P_2^{(i)}, \ldots, P_k^{(i)}\}$.

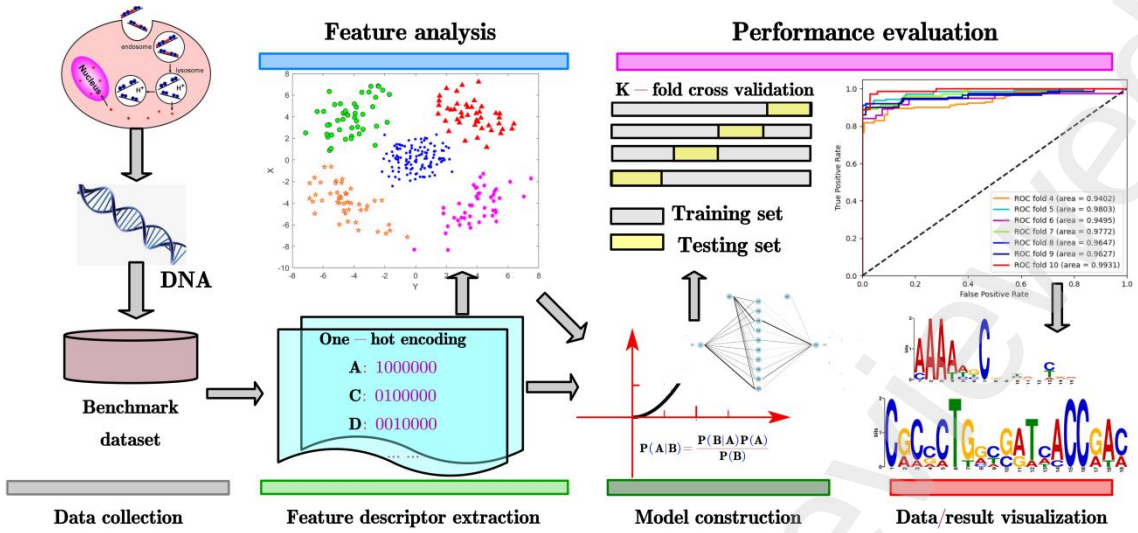


Fig. 3. A summary of several major steps in our work, including data collection, feature extraction, feature analysis, model construction, performance evaluation and results visualization.

3. Experiments and Results

3.1 Implementation Details

In our experiment, the optimizer is ADAM, with the learning rate set to 0.0005. The batch size set to 10 and the iteration step set to 70, respectively. All the experiments were conducted on a computer with NVIDIA GeForce GTX1080Ti graphics card and Intel i7-8700 CPU. The model was implemented using the PyTorch framework and Python language. With 60% of the dataset regarded as the training set, the remaining data were treated as the testing set.

3.2 Evaluation Metrics

Herein, there are six commonly used evaluation metrics applied, including F1-Score (F1), Matthews correlation coefficient (MCC), Sensitivity (Sn), Specificity (Sp), Area under ROC curve (AUC) and Area under PR curve (AUPRC). These metrics can be defined as follows:

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

$$S_n = \frac{TP}{TP + FN}$$

$$S_p = \frac{TN}{TN + FP}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

where *TN* represents true negative, *FN* denotes false negative, *TP* indicates true positive, and *FP* refers to false positive, respectively.

In addition to these evaluation indexes, ROC curve (subject working characteristic curve) is also involved to measure the relationship between true positive rate (TPR) and false positive rate (FPR). In the ROC curve, the real class rate is taken as the vertical axis and the false positive class rate is treated as the horizontal axis. Thus, the calculation formulas are expressed as follows:

$$TPR = \frac{TP}{TP + FN} = recall$$

$$FPR = \frac{FP}{FP + TN}$$

In the ROC curve, the area is enclosed by the curve and the coordinate axis is recorded as AUC (the area under the curve). Its value ranges from 0 to 1. When the value reaches 0.5, it is equivalent to random guess, and the model is of no practical significance. AUC value is positively correlated with model performance.

PR curve is referred to as the curve drawn with precision as the ordinate and recall as the abscissa. The PR curve reveals how precision rate is related to recall rate. The precision rate is defined as the proportion of the real case in all the data predicted as positive cases, and the recall rate refers to the proportion of the data predicted as real cases in all the positive case data. Precision and recall comprise a set of contradictory measures. In generally, the recall is often low when precision is high, while the precision tends to be low when the recall is high. Therefore, the value at the balance point can be used for comparison. As the value obtained when the precision and recall are equal, the balance point takes into account the performance of both aspects. If this value is large, the performance of the model can be improved.

3.3 Evaluation

3.3.1 Some machine learning algorithms

A. Support Vector Machine (SVM)

Proposed by Cortes *et al*. in 1995 [36], support vector machine (SVM) has various unique advantages in processing small sample, nonlinear and high-dimensional pattern

recognition, for which it has been widely adopted to deal with bioinformatics problems. The basic idea of it is to convert the input data into a high-dimensional feature space, before determining the best separation hyper plane, as the decision boundary. Classed as a machine learning method, SVM is based on the statistical learning theory with excellent generalization performance.

The SVM algorithm involves two important parameters: C and γ. Of them, C represents the penalty coefficient, which is the tolerance to the error. Besides, the higher the C, the easier it is to be over-fitted; As a parameter of the function after the selection of the Radial Basis Function (RBF) as the kernel function, γ determines the data after the data is mapped into the new feature space distributed. The larger the γ, the smaller the number of support vectors. Moreover, the speed of training and recognition process can be affected by the number of support vectors.

B. Naïve Bayes (NB)

Assuming that the features are independent of each other [37], Naive Bayes (NB) classifier is a relatively stable supervised classification algorithm. It is premised on the theory of Bayes and has been widely applied to data mining and pattern recognition.

The principle followed by the classification by Bayesian classifier is that the prior probability of an object is used to calculate its posterior probability through the Bayesian formula, that is, the probability that the object belongs to a certain class. Additionally, the class with the largest posterior probability is treated as the class to which it belongs.

C. Logistic Regression (LR)

Known as a classification model in traditional machine learning, Logistic regression (LR) is often performed to deal with binary classification tasks [38]. Currently, it has been widely recognized for such advantages as simplicity, parallelization, and strong interpretability. Besides, it is frequently involved in the second-level base classifier for the stacking integration algorithm. The core data of logistic regression is to assume that the data conform to this distribution, so that maximum likelihood estimation can be relied on to estimate the parameters.

D. Ensemble Learning (EL)

The aim of supervised learning in the context of machine learning is to train a model that shows stability and performs well in all fields. In practice, it is often the case that only the

multiple models with preferences are obtainable, that is, weakly supervised models. The integrated learning model is what consolidates the respective advantages of multiple learning models. By integrating multiple "good but different" models, the correct result can be obtained with the assistance of other weakly supervised models, even in the event of a wrong recognition by one weakly supervised model [39].

The Boosting method combines homogeneous weakly supervised models as well. Different from other methods, however, each model in Boosting depends on the results obtained in the previous round of model learning, and weakly supervised models cease to be independent of each other. One of the typical algorithms is AdaBoost.

E. Random Forest (RF)

As an integrated classification algorithm, the random forest classifier regards a decision tree as the basic unit, the output of which is jointly determined by the output of multiple decision trees. In 1996, Breiman *et al*. proposed the Bagging learning method. According to this method, the training sample set is sampled with replacement to generate a large number of slightly different training sample subsets, and then each sample subset is used to obtain a weak classifier through training [40]. Finally, a more reliable strong classifier can be obtained by integrating weak classifiers, the performance of which is superior to a single weak classifier. Through a combination of the Bagging method and the random subspace division strategy, Bremain *et al*. put forward a random forest (RF) algorithm [41]. In this algorithm, the Bootstrap sampling method is applied to increase the randomness of the data used to construct the classification regression tree. On the basis, multiple classification regression trees are obtained to obtain the final result through prediction.

The RF algorithm demonstrates its advantage in various aspects. Due to the randomness introduced into the algorithm, it is made more robust and less prone to over-fitting, which is conductive to parallel computing. In addition, not only is the RF algorithm capable to classify data, it is also effective in evaluating the importance of variables and analyzing the magnitude of the role placed by each variable in classification [42].

F. Convolutional Neural Network (CNN)

Among the representative algorithms of deep learning, CNN is widely applied in image processing and speech recognition due to its high efficiency of classification [43]. As artificial intelligence technology advances, CNN has been increasingly adopted for the processing of biological information [44].

In general, CNN consists of a convolutional layer, a pooling layer, and a fully connected layer. The convolutional layer and the pooling layer are intended to map the original data into the feature space, while the fully connected layer is purposed to map the learned distributed features into the sample label space and extract important features from DNA sequences automatically, for which it can be regarded as a powerful feature extractor.

3.3.2 Baseline Models

Herein, there are two comparison methods applied, which are traditional machine learning method and deep learning method. Table 2 shows the methods currently available for the prediction of DNA 6mA sites. In the table, there is an in-depth description of the algorithms involved in the methods, testing methods and extracted features.

Table 2 Currently available methods for the prediction of DNA 6mA sites.

Model	Algorithm	Features	Testing Method
Meta-i6mA	SVM, NB, LR, ERT, RF and AdaBoost	TNC, NAC, MBE, Kmer, NCP and EIIP	independent test and 10-fold CV
pm6A-CNN	CNN	One-hot and NCP	10-fold CV
iDNA6mA-PseKNC	SVM, LR, RF, BAG and GNB	RFHC, EIIP, TPCP, DPCP and BE	5-fold CV

3.3.3 Comparison of Results

In most cases, the method of predicting DNA m6A sites is premised on different models with various predefined features. It is believed that the main determinant comes from the feature extraction method although the accuracy of the prediction model can be affected by different algorithm. Therefore, we conducted many experiments to evaluate the impact of different folds on the performance of the model. At the same time, we chose three commonly used methods in the currently published 6mA sites prediction model as the baseline models, namely, Meta-i6mA, pm6A-CNN and iDNA6mA-PseKNC. Upon a comparison of the above three methods against the proposed method, the experimental results were obtained, as shown in Table 3.

Table 3 Comparison of GLF6Ma, Meta-i6mA, pm6A-CNN and iDNA6mA-PseKNC based on different public datasets, in terms of (a) Rice, (b) F. vesca, (c) E. coli and (d) A. thaliana.

(a) Rice

Method	Sn	Sp	ACC	MCC	AUC
Meta-i6mA	0.8410	0.9140	0.8750	0.7520	0.9340
pm6A-CNN	0.8420	0.8800	0.8610	0.7230	0.9230
iDNA6mA-PseKNC	0.5690	0.7210	0.6410	0.3940	0.8960
GLF6mA	**0.9886**	0.8636	**0.9261**	**0.8590**	**0.9879**

(b) F. vesca

Method	Sn	Sp	ACC	MCC	AUC
Meta-i6mA	0.9080	0.9570	0.9340	0.8730	0.9810
pm6A-CNN	0.9781	0.9814	0.9797	0.9590	0.9950
iDNA6mA-PseKNC	0.9200	0.8400	0.9442	0.9300	0.9640
GLF6mA	0.9698	**0.9953**	**0.9826**	**0.9654**	**0.9987**

(c) E. coli

Method	Sn	Sp	ACC	MCC	AUC
Meta-i6mA	0.8580	0.8070	0.8330	0.6660	0.8800
pm6A-CNN	0.8680	0.8270	0.8480	0.6970	0.9000
iDNA6mA-PseKNC	0.8812	0.7888	0.8590	0.6870	0.9210
GLF6mA	**0.8684**	**1.0000**	**0.9342**	**0.8760**	**0.9584**

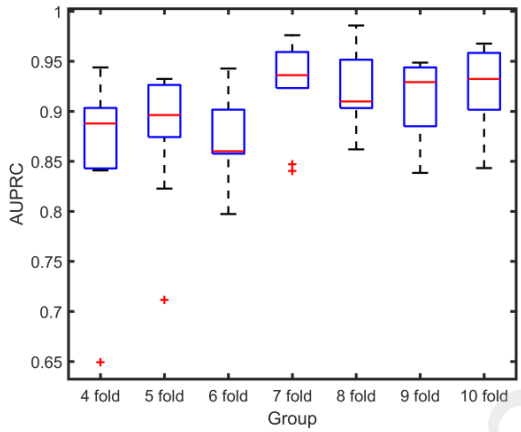
(d) A. thaliana

Method	Sn	Sp	ACC	MCC	AUC
Meta-i6mA	0.9380	1.0000	0.9690	0.9400	0.9710
pm6A-CNN	0.9320	1.0000	0.9660	0.9350	0.9740
iDNA6mA-PseKNC	0.8690	1.0000	0.9350	0.8770	0.9740
GLF6mA	**0.9583**	0.9722	**0.9693**	**0.9366**	**0.9931**

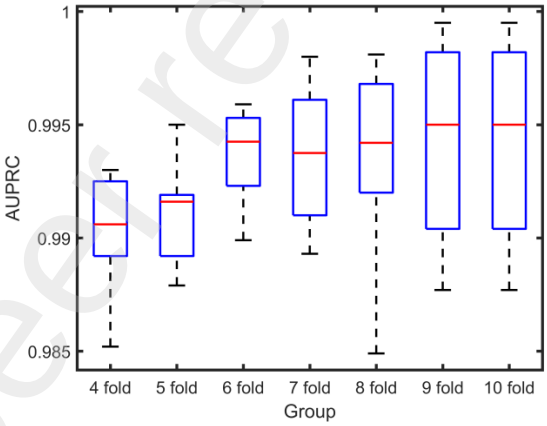
According to the experimental results shown in Table 3, it can be concluded that compared with the model based on the machine learning method, there is a significant improvement of the prediction result produced by the fusion model based on the graph neural network (GLF6mA). Besides, the performance is consistent on these four public datasets. Despite the excellent performance on some datasets, some machine learning methods perform poorly on other datasets, which is a testament to the better adaptability of the fusion model based on GNN. In comparison with Meta-i6mA, pm6A-CNN and iDNA6mA-PseKNC, the level of accuracy is 5.11%, 6.51% and 28.50% higher, respectively. In addition, it also achieves a higher accuracy on the E. coli dataset. However, the experimental result is slightly less satisfactory than on the Rice dataset, with an accuracy of 93.42%, which is 10.12% higher than Meta-i6mA, 8.62% higher than pm6A-CNN, and 7.52% higher than iDNA6mA-PseKNC; GLF6mA is on the A. thaliana dataset. Comparing with Meta-i6mA, pm6A-CNN and iDNA6mA-PseKNC, the level of accuracy is increased by merely 0.03%, 0.33% and 3.43%, despite its accuracy reaching as high as 96.93%. As suggested by these results, GLF6mA can maintain a high level of accuracy on the other three datasets on which no excellent performance is produced, which confirms that the 6mA identification method

based on the GNN-based fusion model as proposed in this study is fit for cross-species identification.

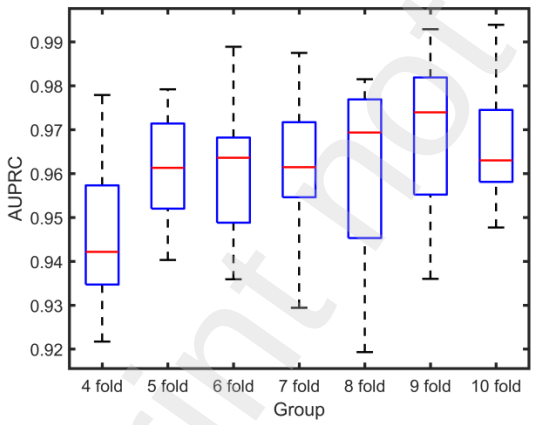
As for the mature DNA, cross-validation testing and independent dataset testing revealed that our improved method outperforms the other three methods in sensitivity, accuracy, AUC and other metrics. For example, in the E. coli dataset, the prediction performance of Meta-i6mA, pm6A-CNN and iDNA6mA-PseKNC was Sp < 0.90, ACC < 90%, AUC < 0.90, while the GLF6mA achieved a better prediction performance of Sp = 1, ACC = 93.42%, AUC = 0.9584, indicating the better performance of GLF6mA in extracting the features of 6mA sites in DNA sequence.



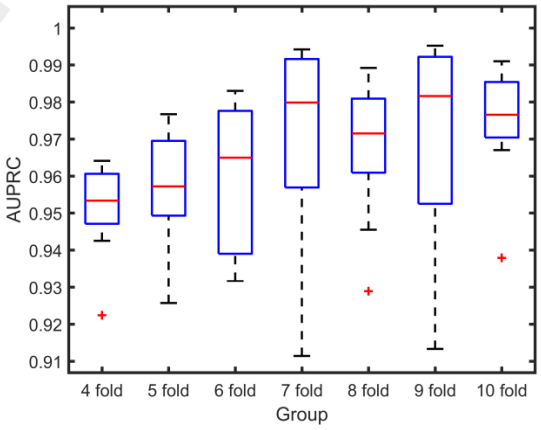
(a) E. coli



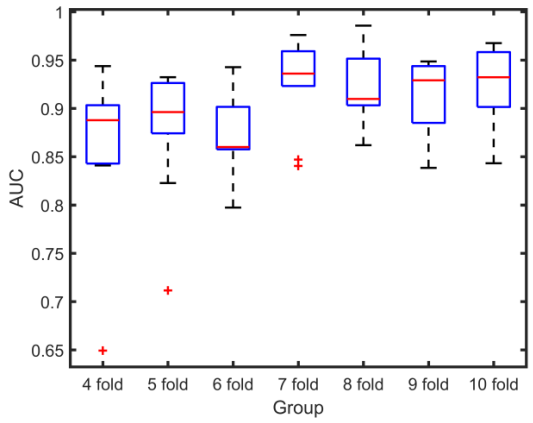
(b) F. vesca



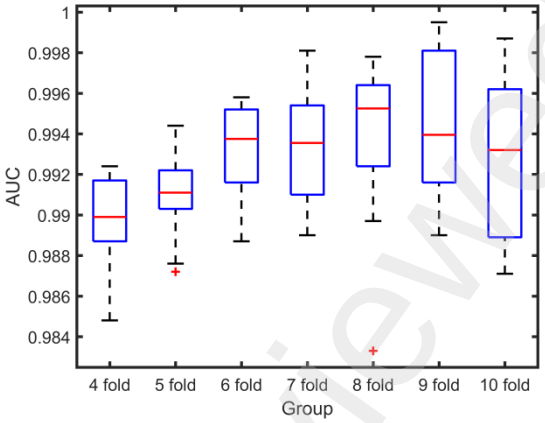
(c) A. thaliana



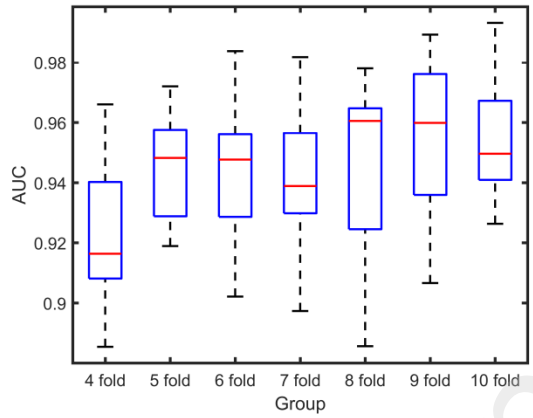
(d) Rice



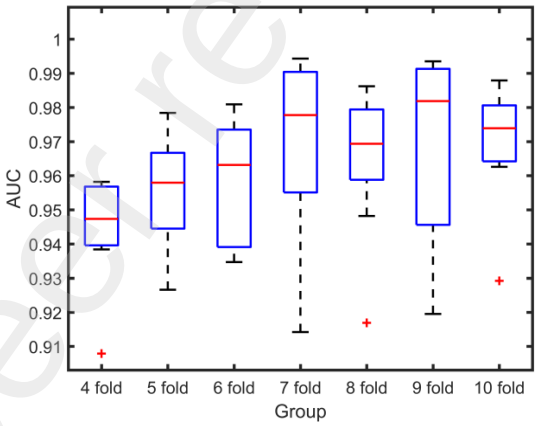
(a) E. coli



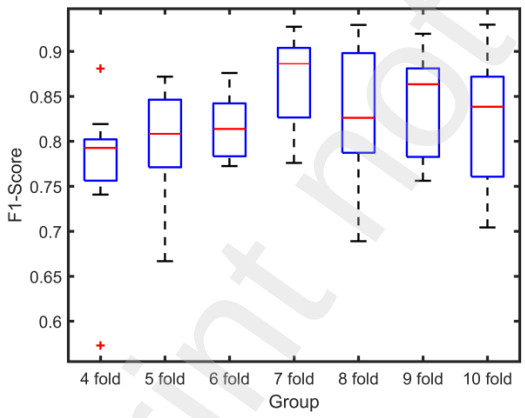
(b) F. vesca



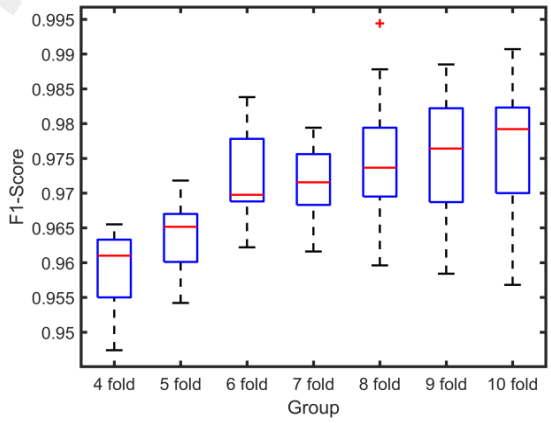
(c) A. thaliana



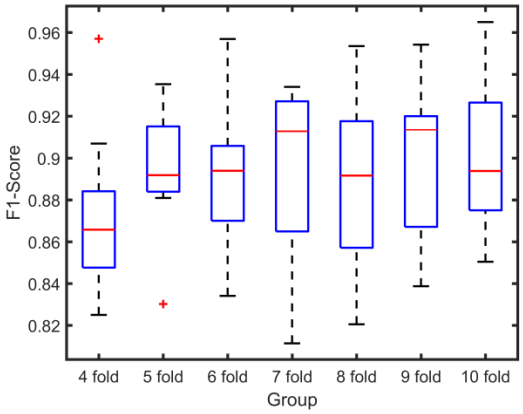
(d) Rice



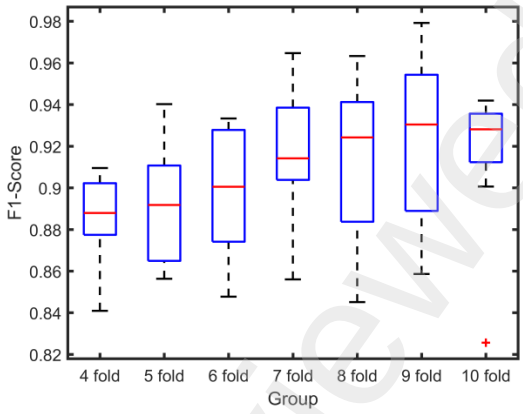
(a) E. coli



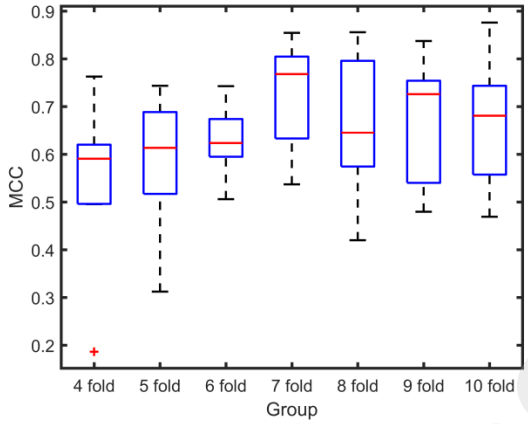
(b) F. vesca



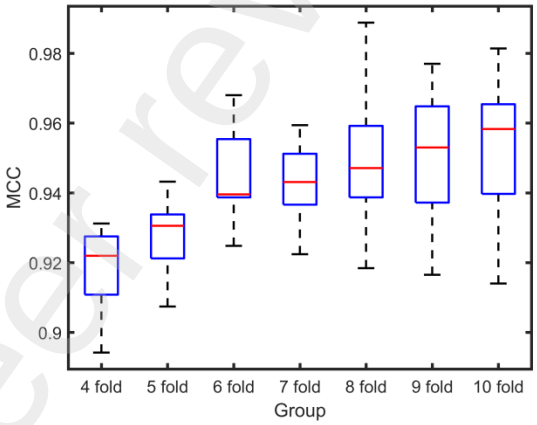
(c) A. thaliana



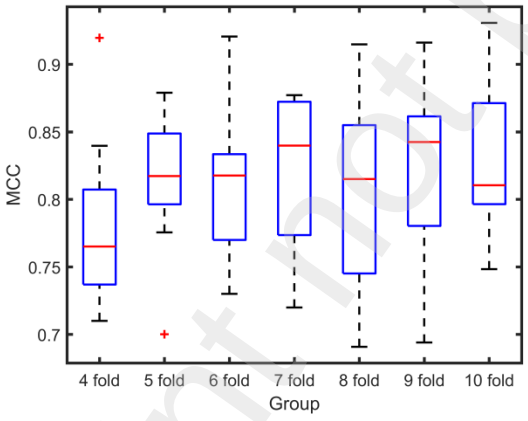
(d) Rice



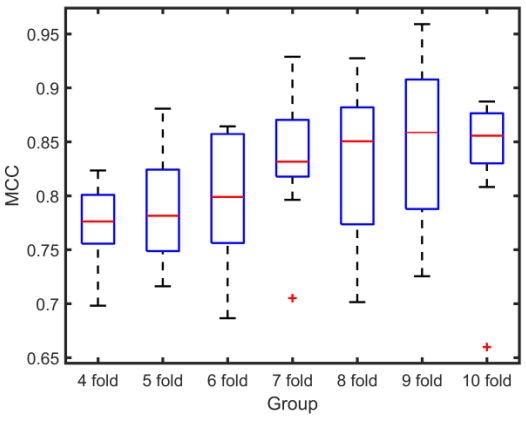
(a) E. coli



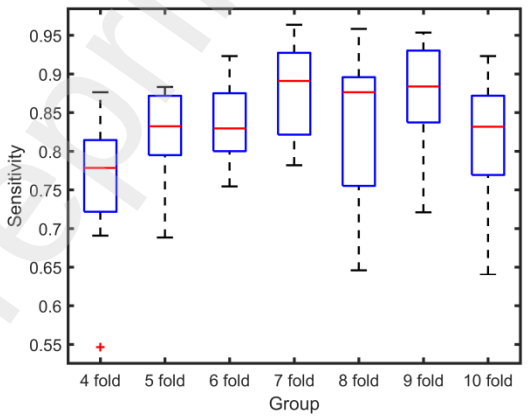
(b) F. vesca

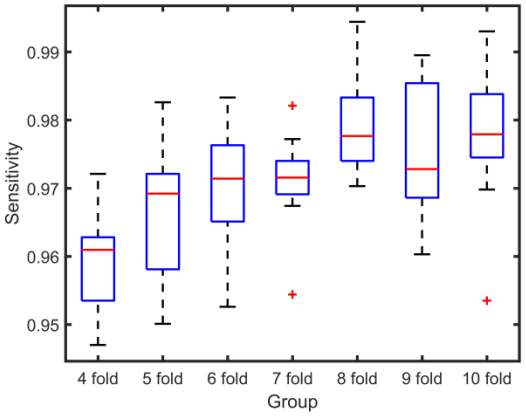


(c) A. thaliana



(d) Rice





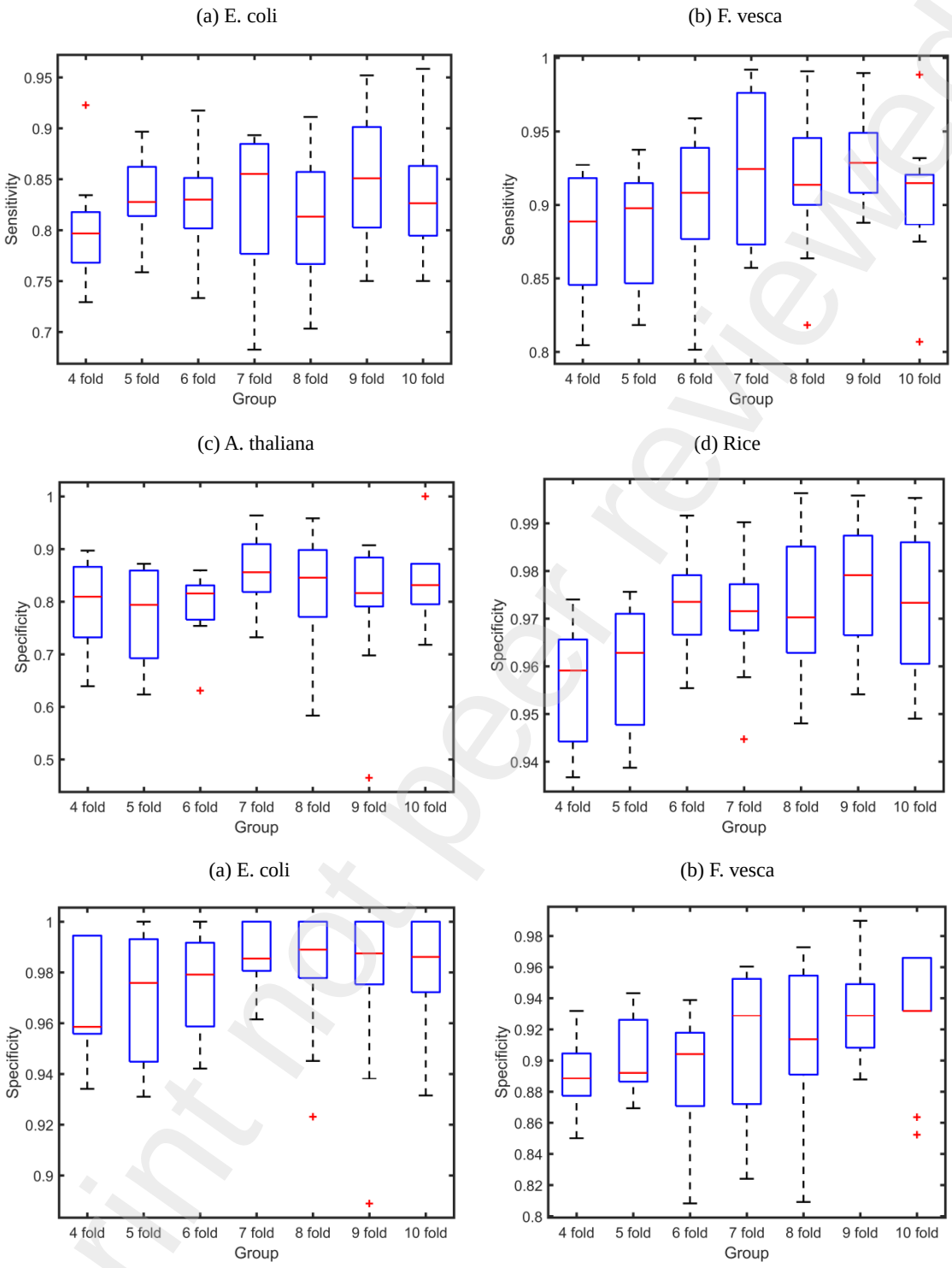
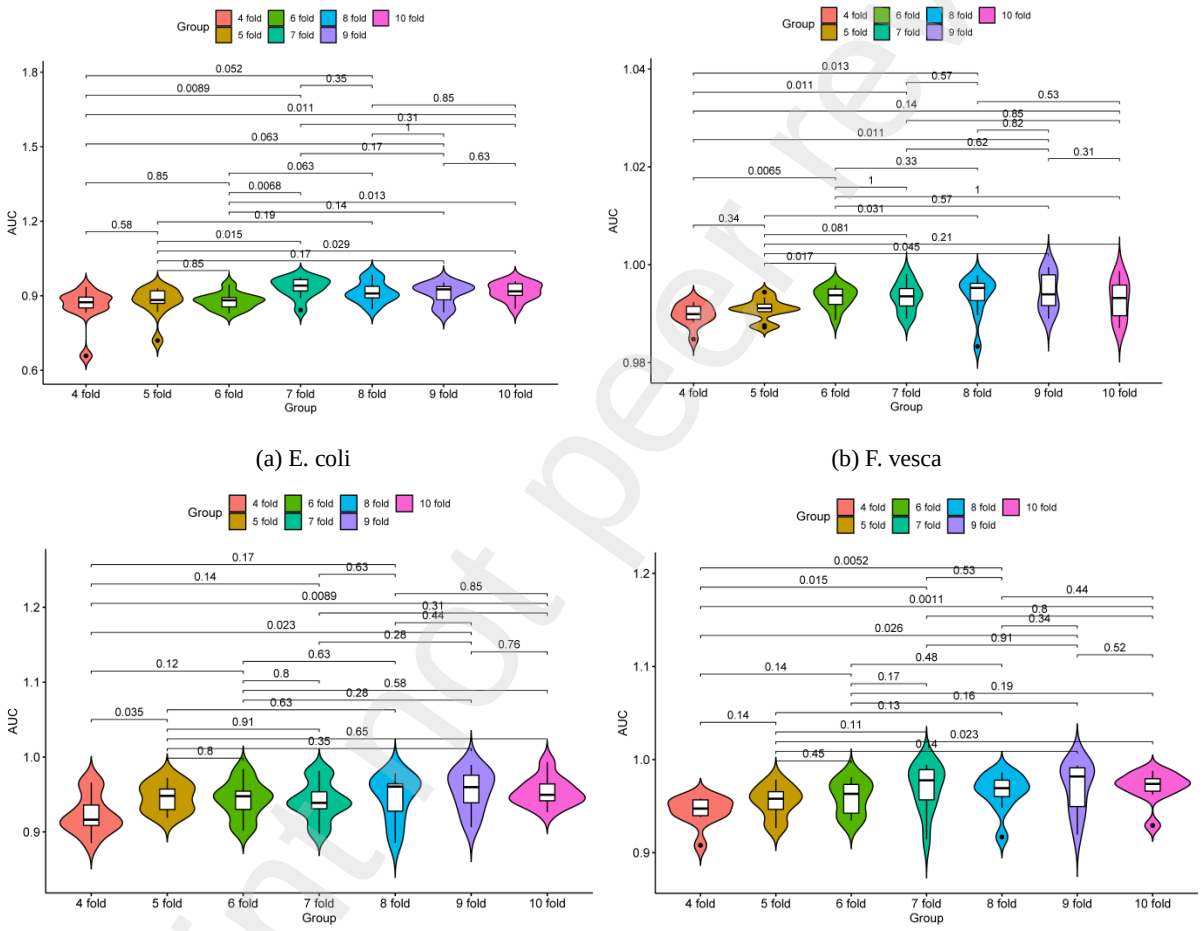


Fig. 4. An illustration of the experimental results generated by GLF6mA for identification DNA 6mA sites in four public datasets, including (a) E. coli, (b) F. vesca, (c) A. thaliana and (d) Rice in terms of six evaluation metrics (A) AUPRC, (B) AUC, (C) F1-Score, (D) MCC, (E) Sensitivity and (F) Specificity.

Furthermore, we evaluated the performance of different models against the ROC curve and PR curve. The AUC is regarded as an important metric for the performance of the classifier. Fig. 6 shows the ROC curves of various prediction models on the four datasets. As

shown in the figure, the AUC value of our proposed model is higher compared to a single classifier model. In contrast, the performance of our proposed model is better. The conclusion drawn from the ROC curve is consistent with that reached on the basis of accuracy comparison in the above table, which substantiates the effectiveness of the fusion model based on GNN. Due to the category imbalance, it is possible for the ROC curve to present an overly optimistic estimate of the effect, which is not encountered by the PR curve. Therefore, the PR curve was used to further evaluate each model for their performance. Fig. 7 shows the PR curves of various models on the four public datasets.



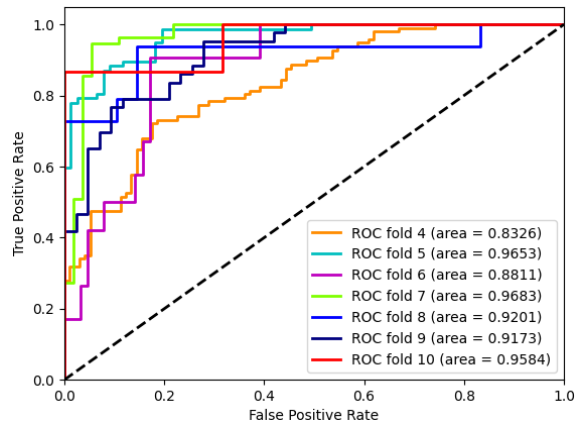
(a) E. coli

(b) F. vesca

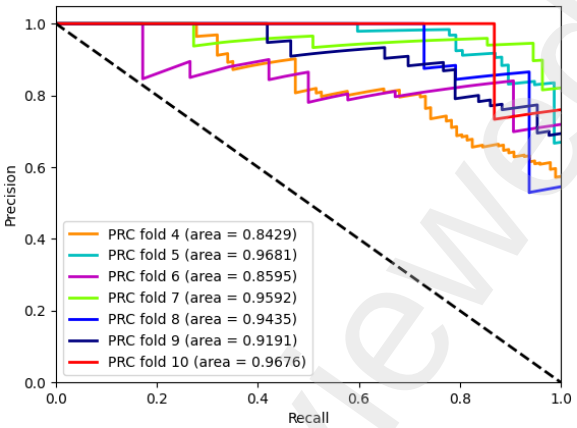
(c) A. thaliana

(d) Rice

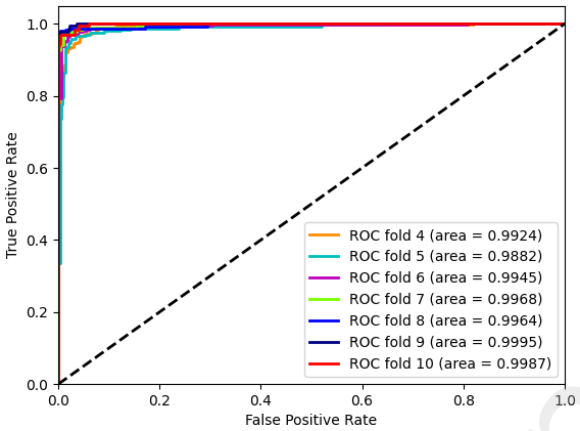
Fig. 5. Performance comparison of GLF6mA on four different public datasets. For each dataset, the AUC values represented no statistical difference ($p > 0.05$), where p is calculated by the student's T-test.



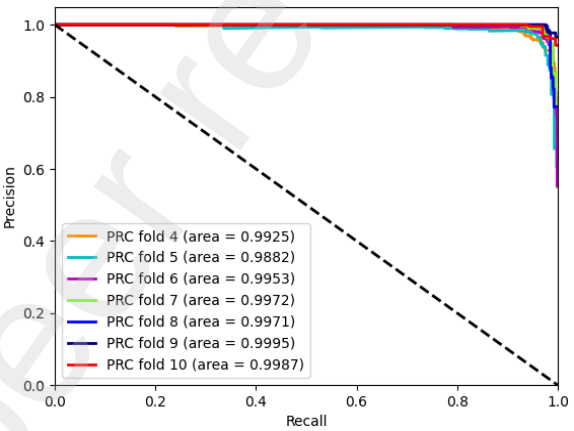
(a) E. coli



(b) F. vesca

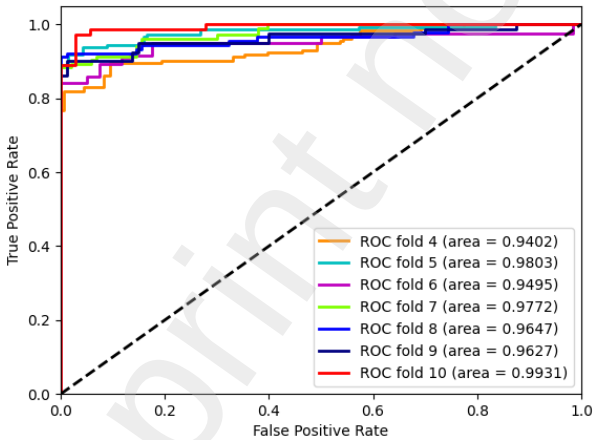


(c) A. thaliana

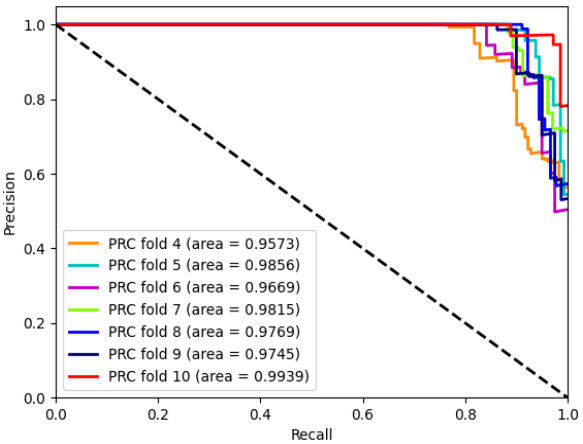


(d) Rice

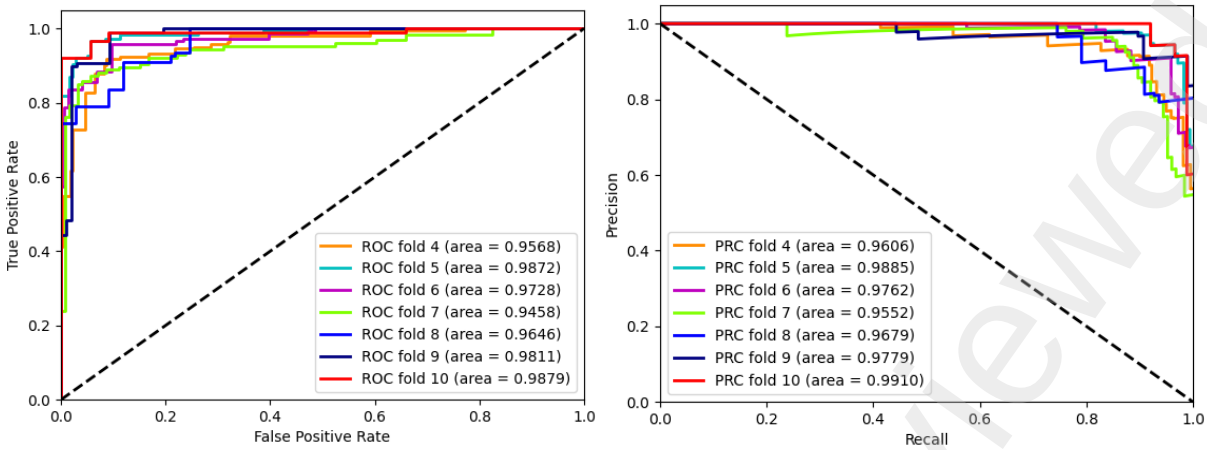
Fig. 6. The ROC curves of GLF6mA on four public datasets, including (a) E. coli, (b) F. vesca, (c) A. thaliana and (d) Rice in terms of seven different folds (A) fold 4, (B) fold 5, (C) fold 6, (D) fold 7, (E) fold 8, (F) fold 9 and (G) fold 10.



(a) E. coli



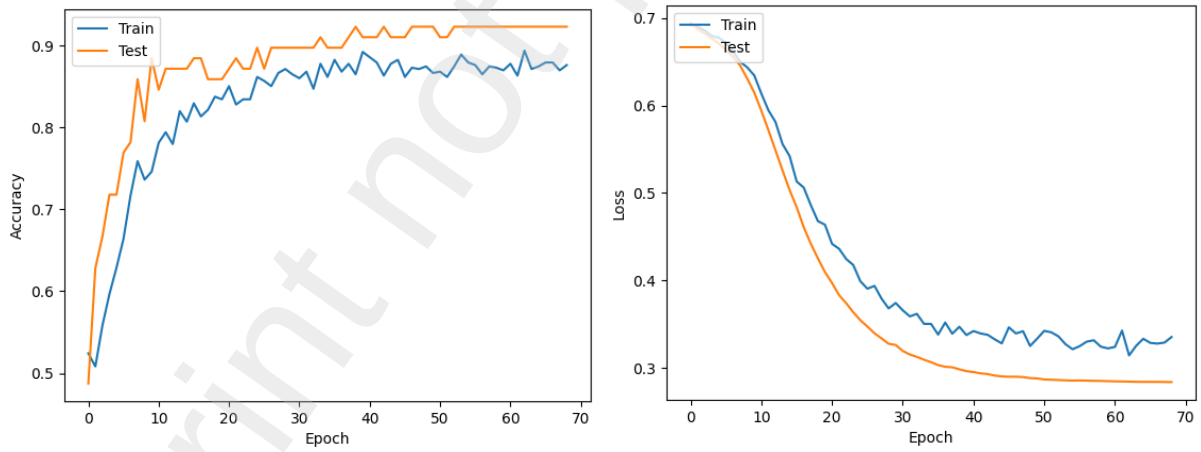
(b) F. vesca



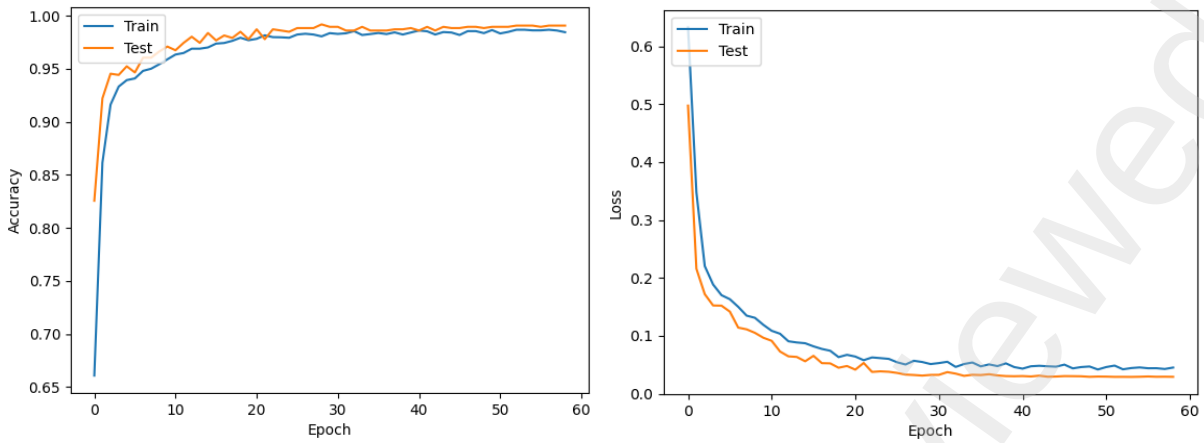
(c) A. thaliana (d) Rice

Fig. 7. The PR curves of GLF6mA on four public datasets, including (a) E. coli, (b) F. vesca, (c) A. thaliana and (d) Rice in terms of seven different folds (A) fold 4, (B) fold 5, (C) fold 6, (D) fold 7, (E) fold 8, (F) fold 9 and (G) fold 10.

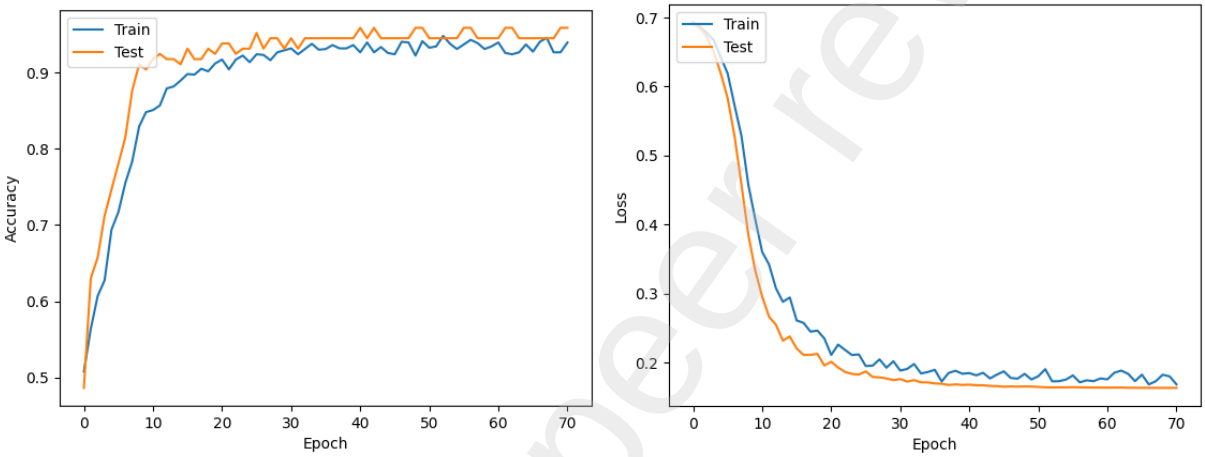
Fig. 8 illustrates the experiment based on the training set and the testing set through our proposed model. According to the change curve, the accuracy of the training set before and after the 50th round reached about 97% prior to convergence. The accuracy of the testing set fluctuated significantly in the early stage, which is because the model did not learned the appropriate parameters well in the first few rounds of model training. After the 20th round, the accuracy of the testing set improved gradually, and converged progressively after the 30th round.



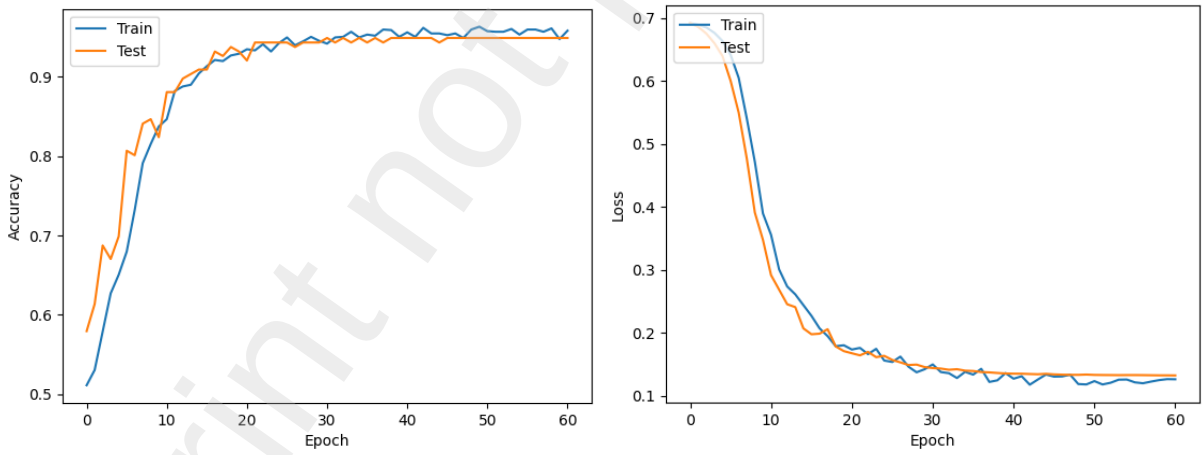
(a) E. coli



(b) F. vesca



(c) A. thaliana



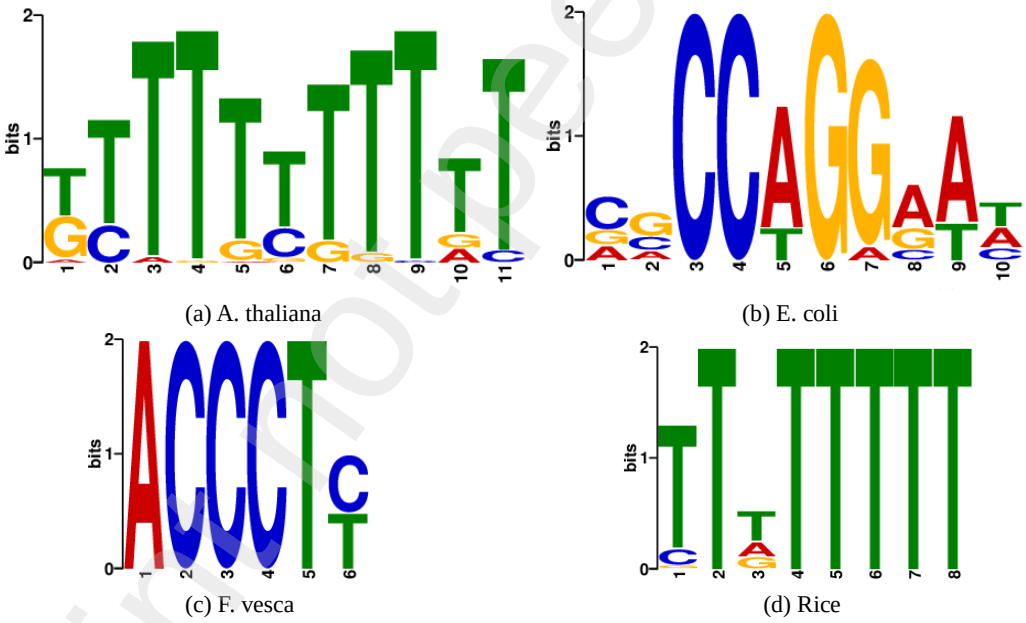
(d) Rice

Fig. 8. The training curves of GLF6mA on four public datasets, including (a) E. coli, (b) F. vesca, (c) A. thaliana and (d) Rice in terms of accuracy and loss.

As shown in Fig. 9 to Fig. 11, the positional features of 6mA sites were studied as well, that is, whether 6mA methylation sites are enriched in specific regions of the genome. Fig. 10 shows the global sequence motifs as extracted by our proposed method on four public datasets. As indicated by the experimental results, the distance distribution between adjacent

6mA sites is similar for different chromosomes, and GAGG is the most significant relevant gene sequence in the four species. Therefore, it can be inferred that DNA 6mA methylation occurs most commonly on the GAGG sequence fragments of different species.

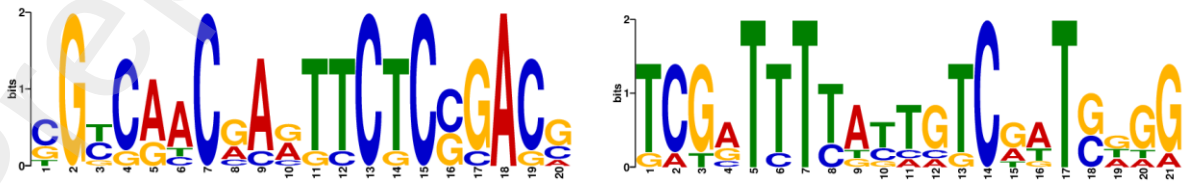
In conclusion, compared with the other three methods, GLF6mA performs better on the independent testing sets of three species, indicating that the GCN based model is more capable of generalization and migration, which makes it fit for the 6mA site identification of multiple species. Since the artificial extraction of DNA features relies heavily on a priori knowledge, it is necessary for feature designers to have solid background knowledge, but the capability of model for cross species 6mA identification is usually insufficient. According to the results of cross species tests on F. vesca, Rice, A. thaliana and E. coli datasets, the GCN based 6mA site identification method GLF6mA shows robustness, the performance of which is better than the three existing 6mA site identification methods and integrated machine learning models, which makes it more suitable for DNA 6mA sites identification cross different species.



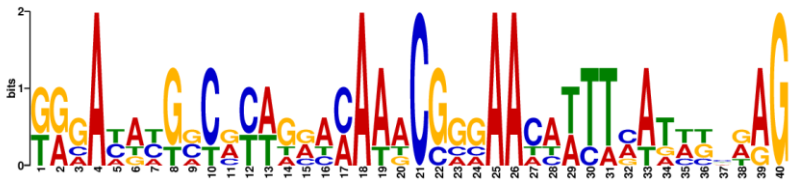
(a) A. thaliana (b) E. coli

(c) F. vesca (d) Rice

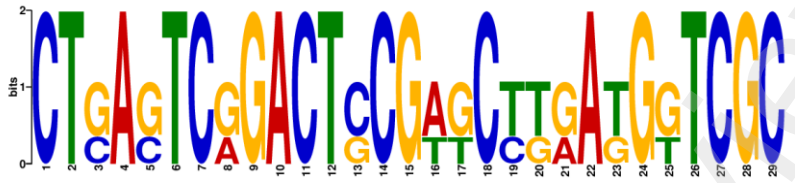
Fig. 9. Local sequence motifs extracted by GLF6mA on four public datasets, including (a) A. thaliana, (b) E. coli, (c) F. vesca and (d) Rice. As for each dataset, we show the most significant mappings of the local sequence motifs extracted by GLF6mA.



(a) A. thaliana (b) E. coli

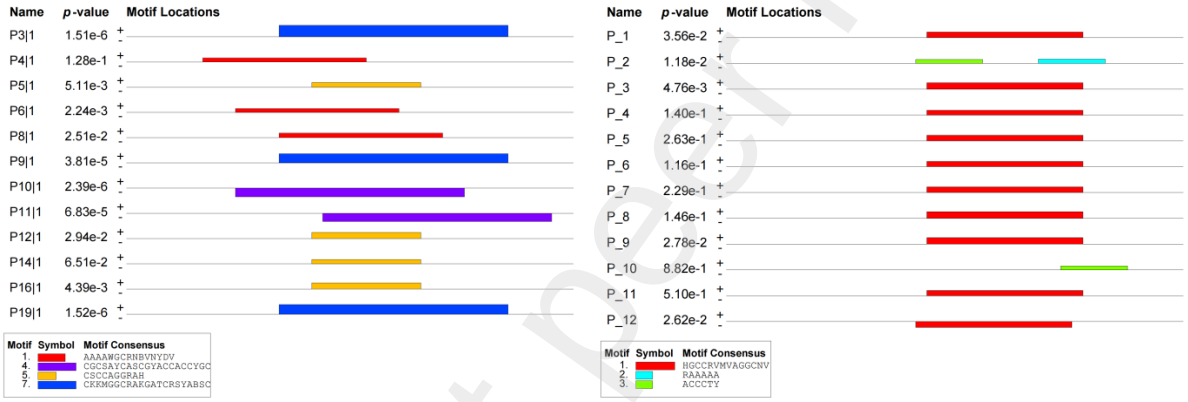


(c) F. vesca



(d) Rice

Fig. 10. Global sequence motifs extracted by GLF6mA on four public datasets, including (a) A. thaliana, (b) E. coli, (c) F. vesca and (d) Rice.



(a) E. coli (b) F. vesca

Fig. 11. Motif locations extracted by GLF6mA on (a) E. coli and (b) F. vesca datasets, where *p*-value is calculated by the student's T-test.

4. Discussion

6mA plays a vital role in maintaining the normal transcriptional activity of cells, repairing DNA damage, chromatin remodeling, genetic imprinting, embryonic development and tumorigenesis [45]. Herein, although deep CNN show its superiority in feature extraction for DNA methylation sequence data, the direct design of large-scale CNN is likely to cause over-fitting due to the high sample dimensions and small number of samples in clinical methylation datasets. Therefore, we constructed a model that could identify DNA 6mA sites by combining GNN, LSTM and FCN. The model was then verified on four public datasets of E. coli, F. vesca, A. thaliana and Rice.

Compared with some other published works, this paper achieves the following improvements. The first one is the *Improvement of the model*. Despite a success of the traditional deep learning method in extracting the characteristics of Euclidean spatial data, the

data are often generated from non-Euclidean space in practice, and the performance of the traditional deep learning method in processing non-Euclidean spatial data remains unsatisfactory. In recent years, there has been attention drawn to the application of deep learning method in graph structure data. Therefore, we constructed a model capable to predict DNA methylation sites by integrating GCN, LSTM and FCN, with experiments performed on multiple public datasets. As suggested by the experimental results, the proposed model performed better in the prediction of DNA methylation sites, with the performance on several metrics improved. The feature extracted before FCN is neither sufficient nor possible enough [9]. The second one is *Experiments*. We conducted a large number of experiments on four datasets, E. coli, F. vesca, A. thaliana and Rice, based on which a large number of experimental data were collected to verify the effectiveness of the model. The four datasets are public, taking into account plants and microorganisms to validate the model and ensure that the conclusions are more universal. In addition, we used several evaluation metrics, including F1-Score, MCC, specificity, sensitivity, AUC, ROC curve, PR curve and Accuracy, etc., so as to evaluate the performance of the proposed model more comprehensively. Moreover, we also adopted three methods as baseline models, including traditional machine learning methods and deep learning methods (CNN), which makes the results more convincing. Finally, study was conducted as to the impact of different folding numbers on the outcome of model training with the experimental results verified from fold 4 to fold 10. In the previously published papers, the author only used AUC as the evaluation metric, which is a limitation. Though the results may be better on AUC, the effect is average when other metrics are used for evaluation. Besides, there were no other models available for comparison. As a result, only their own method was used to carry out experiments on some datasets, with only three datasets used by the authors [9]. The third one is *Performance*. The model designed by us has achieved satisfactory results. Our effect is better compared to the baseline models whether in the verification of individual different discounts or the overall average value. When the same dataset was used as in the published paper, we compared the average value of AUC under the same conditions, which led to the finding that the effect of our proposed model is also better than the experimental results obtained in this study. On the A. thaliana and E. coli datasets, our AUC averages were 0.9931 and 0.9584, respectively, suggesting that our proposed model is more capable of generalization and migration, which makes it fit for identifying and predicting 6mA sites of multiple species. In previously published papers, the average AUC values corresponding to the two datasets were 0.9564 and 0.9390 respectively [9]. The last one is *Analysis*. We also studied the positional characteristics of 6mA sites, that

is, whether 6mA methylation sites are enriched in specific regions of the genome. A study was conducted on the global and local sequence motifs as extracted by our proposed method on four public datasets. According to the experimental results, the distance distribution between adjacent 6mA site is similar for different chromosomes, and GAGG is the most important relevant gene sequence in the four species. Therefore, it can be inferred that DNA 6mA methylation most often occurs on the GAGG sequence fragments of different species. In the previously published papers, the authors studied the global sequence motifs without inference and analysis, which leads to certain limitations. Moreover, 6mA methylation is demonstrated by some researches as a dynamic regulation process, and the methylation state may change with different stages of biological development or tissue-specific changes. It is thus necessary to study the effects of these factors [9].

Despite the better results produced by the model proposed by us for DNA methylation modification site prediction on multiple DNA sequence datasets, there remain some problems. Firstly, given the discovery of 6mA sequences of more species, we will collect 6mA sequences of more species in the future, so as to further improve the performance of the model. Moreover, the current practice of identifying 6mA sites is mostly a binary classification problem, which means it is practical to study a multi-classification model that can identify 6mA sites of multiple species at the same time. Secondly, there is a lack of combination of RNA sequence data and DNA methylation data for simultaneous prediction. RNA-Sequence data and DNA methylation data represent two different types of data, which can reflect the characteristic information of the sample from different angles. Besides, the combination of data can provide a more reliable basis for classification. However, it is not suitable for the training of machine learning models because there are both RNA-Sequence data and DNA methylation data in the current database and the number of samples is relatively small, especially the normal type of samples. In the future, we plan to improve the universality of the model for modeling more types of DNA methylation data.

5. Conclusion

As for DNA methylation data, despite the advantages of deep neural networks in feature extraction, it is easy to directly design large-scale CNN due to the high sample feature dimension and small number of samples in DNA methylation datasets. In this regard, we proposed a model (GLF6mA) that could combine GNN, LSTM and FCN to predict DNA methylation sites, before experiments were conducted on multiple datasets. According to the experimental results, our proposed model performed better in the prediction of DNA

methylation site, with a significant improvement achieved for many performance metrics. Moreover, the deep learning prediction model is highly sensitive to the size of the dataset. For large datasets, its prediction performance is even better. As the amount of training set decreases, however, its prediction performance will deteriorate. In the future, we will construct a 6mA site dataset with a larger amount of data involved and develop a set of deep learning prediction models appropriate for small datasets. In addition, we plan to construct a standard dataset related to diseases, clarify the correlation between diseases and 6mA sites, and improve the scientific basis for medical treatment, drug research and development in the future.

Acknowledgements

The authors gratefully acknowledge the financial supports of the Natural Science Foundation of Fujian Province (2019J01272), the United Fujian Provincial Health and Education Project for Tackling the Key Research of China (2019-WJ-03) and the Special Funds of the Central Government Guiding Local Science and Technology Development (2020L3008).

Declaration of Interests

The authors declare no conflicts of interest.

Author Contributions

Qing Wang, Lin Song, Guannan Chen and Yao Lin are the main authors of this article. Guannan Chen and Yao Lin proposed and deduced the main idea and methodology, Qing Wang and Lin Song completed experiments and wrote this article. Weiping Liu, Xinghong Chen, Xiumei Wang, Bin Yang, Juhao Jiang and Xuran Zhou analyzed and verified experimental results. All authors read and approved the final manuscript.

References

[1] Loughland, Isabella, Alexander Little, and Frank Seebacher. “DNA methyltransferase 3a mediates developmental thermal plasticity.” *BMC Biology* 19.1 (2021): 1-11.

[2] Cescon, David W., *et al*. “Circulating tumor DNA and liquid biopsy in oncology.” *Nature Cancer* 1.3 (2020): 276-290.

[3] Nakamura, Yoshiaki, *et al*. “Clinical utility of circulating tumor DNA sequencing in advanced gastrointestinal cancer: SCRUM-Japan GI-SCREEN and GOZILA studies.” *Nature Medicine* 26.12 (2020): 1859-1864.

[4] Zhao, Shuang G., *et al*. "The DNA methylation landscape of advanced prostate cancer." *Nature Genetics* 52.8 (2020): 778-789.

[5] Koelsche, Christian, *et al*. "Sarcoma classification by DNA methylation profiling." *Nature Communications* 12.1 (2021): 1-10.

[6] Lv, Hao, *et al*. "Advances in mapping the epigenetic modifications of 5-methylcytosine (5mC), N6-methyladenine (6mA), and N4-methylcytosine (4mC)." *Biotechnology and Bioengineering* 118.11 (2021): 4204-4216.

[7] Shen, Lu, *et al*. "Genome-wide analysis of DNA methylation in 106 schizophrenia family trios in Han Chinese." *EBioMedicine* 72 (2021): 103609.

[8] Yang, Ruimeng, *et al*. "The RNA methyltransferase NSUN6 suppresses pancreatic cancer development by regulating cell proliferation." *EBioMedicine* 63 (2021): 103195.

[9] Tan, Fei, *et al*. "Elucidation of DNA methylation on N6-adenine with deep learning." *Nature Machine Intelligence* 2.8 (2020): 466-475.

[10] Chen, Zheng, Yan Liu, and Hailin Wang. "Conjoint expression and purification strategy for acquiring proteins with ultra-low DNA N6-methyladenine backgrounds in Escherichia coli." *Bioscience Reports* 41.3 (2021): BSR20203769.

[11] Gao, Yang, Martin Widschwendter, and Andrew E. Teschendorff. "DNA methylation patterns in normal tissue correlate more strongly with breast cancer status than copy-number variants." *EBioMedicine* 31 (2018): 243-252.

[12] Tarsounas, Madalena, and Patrick Sung. "The antitumorigenic roles of BRCA1–BARD1 in DNA repair and replication." *Nature Reviews Molecular Cell Biology* 21.5 (2020): 284-299.

[13] Gibbs, Richard A. "The human genome project changed everything." *Nature Reviews Genetics* 21.10 (2020): 575-576.

[14] Lorenzo, Maria Paz, *et al*. "Optimization and validation of a chiral CE-LIF method for quantitation of aspartate, glutamate and serine in murine osteocytic and osteoblastic cells." *Journal of Chromatography B* 1152 (2020): 122259.

[15] Mehta, Devang, *et al*. "Full-length sequencing of circular DNA viruses and extrachromosomal circular DNA using CIDER-Seq." *Nature Protocols* 15.5 (2020): 1673-1689.

[16] Wong, Chia-En, *et al*. "TDP-43 proteinopathy impairs mRNP granule mediated postsynaptic translation and mRNA metabolism." *Theranostics* 11.1 (2021): 330.

[17] Xiong, Yuanpeng, *et al*. "Modeling multi-species RNA modification through multi-task curriculum learning." *Nucleic Acids Research* 49.7 (2021): 3719-3734.

[18] Garg, Anjali, *et al*. “mRNALoc: a novel machine-learning based in-silico tool to predict mRNA subcellular localization.” *Nucleic Acids Research* 48.W1 (2020): W239-W243.

[19] Wang, Senzhang, Jiannong Cao, and Philip Yu. “Deep learning for spatio-temporal data mining: A survey.” *IEEE Transactions on Knowledge and Data Engineering* (2020).

[20] Du, Haizhou, and Yan Zhou. “Nostradamus: A novel event propagation prediction approach with spatio-temporal characteristics in non-Euclidean space.” *Neural Networks* 145 (2022): 386-394.

[21] Lee, Yeonkun, *et al*. “SpherePHD: Applying CNNs on 360° Images with Non-Euclidean Spherical PolyHeDron Representation.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[22] Zhou, Ding-Xuan. “Theory of deep convolutional neural networks: Downsampling.” *Neural Networks* 124 (2020): 319-327.

[23] Weerakody, Philip B., *et al*. “A review of irregular time series data handling with gated recurrent neural networks.” *Neurocomputing* 441 (2021): 161-178.

[24] Sayers, Eric W., *et al*. “Database resources of the national center for biotechnology information.” *Nucleic Acids Research* 49.D1 (2021): D10.

[25] Hasan, Md Mehedi, *et al*. “i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation.” *Plant Molecular Biology* 103.1 (2020): 225-234.

[26] Boquete, M. Teresa, Aline Muyle, and Conchita Alonso. “Plant epigenetics: phenotypic and functional diversity beyond the DNA sequence.” *American Journal of Botany* 108.4 (2021): 553-558.

[27] Harley, Calvin B., and Robert P. Reynolds. “Analysis of E. coli pormoter sequences.” *Nucleic Acids Research* 15.5 (1987): 2343-2361.

[28] Hasan, Md Mehedi, *et al*. “Critical evaluation of web-based DNA N6-methyladenine site prediction tools.” *Briefings in Functional Genomics* (2021).

[29] Zügner, Daniel, Amir Akbarnejad, and Stephan Günnemann. “Adversarial attacks on neural networks for graph data.” *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018.

[30] Vanetik, Natalia, Solomon Eyal Shimony, and Ehud Gudes. “Support measures for graph data.” *Data Mining and Knowledge Discovery* 13.2 (2006): 243-260.

[31] Wu, Zonghan, *et al*. “A comprehensive survey on graph neural networks.” *IEEE Transactions on Neural Networks and Learning Systems* 32.1 (2020): 4-24.

[32] Gama, Fernando, Joan Bruna, and Alejandro Ribeiro. “Stability properties of graph

neural networks." *IEEE Transactions on Signal Processing* 68 (2020): 5680-5695.

[33] Abdel-Hamid, Ossama, *et al*. "Convolutional neural networks for speech recognition." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.10 (2014): 1533-1545.

[34] Jiang, Wenjun, *et al*. "Understanding graph-based trust evaluation in online social networks: Methodologies and challenges." *ACM Computing Surveys (Csur)* 49.1 (2016): 1-35.

[35] Zhou, Xiang, *et al*. "Graph convolutional network hashing." *IEEE Transactions on Cybernetics* 50.4 (2018): 1460-1472.

[36] Cortes, Corinna, and Vladimir Vapnik. "Support vector machine." *Machine Learning* 20.3 (1995): 273-297.

[37] Murphy, Kevin P. "Naive bayes classifiers." *University of British Columbia* 18.60 (2006): 1-8.

[38] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression*. Vol. 398. John Wiley & Sons, 2013.

[39] Sagi, Omer, and Lior Rokach. "Ensemble learning: A survey." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.4 (2018): e1249.

[40] Breiman, Leo. "Bagging predictors." *Machine learning* 24.2 (1996): 123-140.

[41] Breiman, Leo. "Random forests." *Machine Learning* 45.1 (2001): 5-32.

[42] Speiser, Jaime Lynn, *et al*. "A comparison of random forest variable selection methods for classification prediction modeling." *Expert Systems with Applications* 134 (2019): 93-101.

[43] Sarıgül, Mehmet, Buse Melis Ozyildirim, and Mutlu Avci. "Differential convolutional neural network." *Neural Networks* 116 (2019): 279-287.

[44] Yan, Xiliang, *et al*. "Prediction of Nano–Bio Interactions through Convolutional Neural Network Analysis of Nanostructure Images." *ACS Sustainable Chemistry & Engineering* 8.51 (2020): 19096-19104.

[45] Chen, Zhen, *et al*. "iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization." *Nucleic Acids Research* 49.10 (2021): e60-e60.