OXFORD

Briefings in Bioinformatics, 2022, 1–15
https://doi.org/10.1093/bib/bbac082
Problem Solving Protocol

MGF6mARice: prediction of DNA N6-methyladenine sites in rice by exploiting molecular graph feature and residual block

Mengya Liu, Zhan-Li Sun, Zhigang Zeng and Kin-Man Lam

Corresponding author. Zhan-Li Sun, School of Electrical Engineering and Automation, Anhui University, Hefei 230601, China. E-mail: zhlsun2006@126.com

Abstract

DNA N6-methyladenine (6mA) is produced by the N6 position of the adenine being methylated, which occurs at the molecular level, and is involved in numerous vital biological processes in the rice genome. Given the shortcomings of biological experiments, researchers have developed many computational methods to predict 6mA sites and achieved good performance. However, the existing methods do not consider the occurrence mechanism of 6mA to extract features from the molecular structure. In this paper, a novel deep learning method is proposed by devising DNA molecular graph feature and residual block structure for 6mA sites prediction in rice, named MGF6mARice. Firstly, the DNA sequence is changed into a simplified molecular input line entry system (SMILES) format, which reflects chemical molecular structure. Secondly, for the molecular structure data, we construct the DNA molecular graph feature based on the principle of graph convolutional network. Then, the residual block is designed to extract higher level, distinguishable features from molecular graph features. Finally, the prediction module is used to obtain the result of whether it is a 6mA site. By means of 10-fold cross-validation, MGF6mARice outperforms the state-of-the-art approaches. Multiple experiments have shown that the molecular graph feature and residual block can promote the performance of MGF6mARice in 6mA prediction. To the best of our knowledge, it is the first time to derive a feature of DNA sequence by considering the chemical molecular structure. We hope that MGF6mARice will be helpful for researchers to analyze 6mA sites in rice.

Keywords: DNA N6-methyladenine, SMILES, DNA molecular graph feature, residual block, rice genome

Introduction

As one kind of epigenetic mechanism, DNA methylation is related to many biological functions [1]. The common types of DNA methylation are N4-methylcytosine (4mC), 5-methylcytosine (5mC) and N6-methyladenine (6mA) [2, 3]. Among these, 6mA is a rising adenine (a nitrogenous base in DNA) modification (N6 position of adenine modified by methyl (-CH_3)) found in higher eukaryotes [4, 5]. In recent years, more and more studies have shown that 6mA in rice plays an important role in many biological functions. For example, 6mA in rice (i) inhibits transcription, which in turn regulates gene expression [6]; (ii) is related to stress response and adapts to adversity [7]; (iii) is relevant to reproduction and adjusts the growth and development of rice [8].

Currently, for plants, the commonly used and accurate technologies for detecting 6mA in biological lab mainly include methylated DNA immunoprecipitation sequencing for 6mA (6mA-IP-seq), 6mA-IP-seq combining photo-crosslinking with exonuclease digestion (6mA-CLIP-exo-seq), methylated DNA with restriction enzyme digestion followed by sequencing (6mA-RE-seq) and single-molecule real-time sequencing (SMRT-seq) [9]. The last two methods can detect 6mA sites at single-base resolution [10, 11]. However, 6mA-RE-seq cannot detect all 6mA sites when there are low 6mA proportions in the genome because of restriction enzymes that require specific sequence context and incomplete digestion [11, 12]. Relatively speaking, compared with the above methods, SMRT-seq is currently a better method in terms of the accuracy and robustness of 6mA detection [13] and has been used for the detection of 6mA in a variety of eukaryotes [14–16]. Nevertheless, SMRT-seq requires high sequencing coverage and high cost, which limits its large-scale applications [14].

Similar to the development process of 4mC and 5mC sites prediction methods [17, 18], due to the shortcomings of the above-mentioned biological experimental

Mengya Liu is a Ph.D. candidate in the School of Computer Science and Technology, Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui University, Hefei, China. Her current research interests include bioinformatics and machine learning.
Zhan-Li Sun is a Professor in the School of Electrical Engineering and Automation, Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui University, Hefei, China. His current research interests include machine learning, and image and signal processing.
Zhigang Zeng is a Professor in the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. His current research interests include theory of functional differential equations and differential equations with discontinuous right-hand sides, and their applications to dynamics of neural networks, memristive systems, and control systems.
Kin-Man Lam is a Professor in the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, China. His current research interests include human face recognition, image and video processing, and computer vision.
Received: November 22, 2021. **Revised:** February 13, 2022. **Accepted:** February 16, 2022

detection methods, researchers have developed many computational prediction methods for 6mA sites in the rice genome nowadays. These methods can be roughly divided into traditional machine learning (ML) methods and deep learning methods, either using a single feature or diverse features, summarized in Table 1.

To be specific, in 2019, by using nucleotide chemical property (NCP) feature and support vector machine (SVM) classifier, Chen *et al.* developed the first method to predict DNA 6mA sites in rice, called i6mA-Pred [19]. Subsequently, more and more prediction methods have been proposed based on a single feature and ML classifier, such as MM-6mAPred [21], i6mA-DNCP [24] and iN6-methylat [20]. Since a single feature cannot extract useful information well, many studies have begun to adopt multiple features to predict 6mA sites, including monomer binary (MB), Kmer, pseudo K-tuple nucleotide composition (PseKNC), electron-ion interaction potential (EIIP), position-specific trinucleotide propensity based on double-strand (PSTNPds) and natural vector [22, 28, 29, 31]. Considering that the ensemble of multiple classifiers can further improve model accuracy and robustness [41], 6mA-Finder [30], csDMA [23], i6mA-VC [35], Meta-i6mA [32] and SDM6A [25] were proposed. With the development of deep learning and its outstanding performance, researchers began to apply it to the DNA 6mA sites prediction problem in rice, such as iDNA6mA [26], iRicem6A-CNN [34], Deep6mA [37], DNA6mA-MINT [33], SNNRice6mA [27] and SMEP [39]. Recently, by integrating traditional features, Rahman *et al.* proposed a deep learning method, i6mA-CNN [38]. In addition to integrating features, 6mAPred-MSFF was developed by using convolutional neural network, bidirectional long short-term memory and multi-scale attention mechanism [40] for predicting 6mA sites in rice.

Even if the existing methods have achieved good performance, 6mA sites prediction in rice still remains a deficiency. As mentioned above, in the molecular structure of adenine, the N6 position is modified by -CH_3, which produces DNA 6mA. However, the existing methods mainly used features based on nucleic acid (pairs) type, frequency of nucleotide (pairs) and physicochemical properties. In other words, no feature considers the molecular level of the DNA 6mA production mechanism and is extracted from the molecular structure of DNA bases. Meanwhile, inspired by successful applications of the molecular structure feature in drug and peptide sequence [42, 43], and the current gradual rise in research about a microscopic view of biochemical molecules [44, 45], a 6mA sites prediction method in rice is proposed according to the chemical molecular structure of DNA bases in this study. One DNA sequence is preliminarily expressed as the simplified molecular input line entry system (SMILES) string, which describes the chemical structure of the molecule [46]. For the graph data composed of chemical molecular structure, the basic principle of graph convolutional network (GCN) [47] is used to construct the new feature, that is, the DNA molecular graph feature (MGF). In addition, considering that the residual block is widely used in bioinformatics [48], it is also utilized in this research, due to its ability to extract higher level and easier-to-distinct features. As far as we know, there are very few applications of the residual block in the 6mA sites prediction of rice.

As the main work of this study, we propose an effective deep learning method based on Molecular Graph Feature and residual block for 6mA prediction in Rice, referred to as MGF6mARice. Compared with the state-of-the-art methods, demonstrated by multi-faceted experiments, MGF6mARice has better performance and comparable robustness. More importantly, through comparative experiments between different features and classifiers, MGF is obviously more suitable and effective for 6mA sites prediction in rice. To the best of our knowledge, this is the first time to devise a feature of DNA sequence in terms of the chemical molecular structure.

Materials and Methods

Datasets

The performance of MGF6mARice is evaluated on three kinds of datasets, i.e. benchmark datasets Rice:Chen and Rice:Lv, imbalanced datasets constructed via Rice:Chen or Rice:Lv, two types of independent datasets.

Benchmark datasets

The dataset Rice:Chen is the first high-quality DNA 6mA benchmark dataset in the rice genome [19]. After obtaining SMRT-seq data from NCBI GEO (Gene Expression Omnibus) via the accession number GSE103145 [15], the positive samples of Rice:Chen were produced according to Methylome Analysis Technical Note. And the negative samples of Rice:Chen were fetched from the sequence that was not methylated proved by experimental results, or from motifs where 6mA was less enriched. Compared with Rice:Chen, Rice:Lv is a larger dataset constructed from iDNA6mA-Rice [22]. The acquisition rules about positive and negative samples of Rice:Lv are similar to those of Rice:Chen. In both datasets, the CD-HIT tool [49] was adopted to reduce homology bias and de-redundancy. Considering the huge time and space cost, the construction and optimization of the model use the small dataset Rice:Chen, and the Rice:Lv is used to evaluate the performance of the model on a large amount of data.

Imbalanced datasets

In nature, the number of negative samples is generally far larger than that of positive samples. Sometimes, sample imbalance may significantly decrease the classification performance. In order to investigate the robustness of various methods to sample imbalance, six imbalanced datasets are constructed. Based on Rice:Chen and Rice:Lv, positive and negative samples of these imbalanced datasets are randomly selected via 1:5, 1:10 and 1:20 of positive to negative samples selection ratio, respectively [38].

Table 1. A statistics of features for DNA 6mA sites prediction in the rice genome

Year	Methods	Feature category	Feature encoding
2019	i6mA-Pred [19]	Physicochemical property	NCP
	iN6-methylat [20]	Natural language processing	Continuous bags of nucleobases
	MM-6mAPred [21]	Physicochemical property	NCP
	iDNA6mA-Rice [22]	Nucleic acid (pairs) type	MB
		Nucleotide (pairs) frequency	PseKNC
		Natural language processing	Natural vector
	csDMA [23]	Nucleotide (pairs) frequency	Kmer, KSNPF, Nucleic shift density
		Physicochemical property	NCP
	i6mA-DNCP [24]	Physicochemical property	DPP
	SDM6A [25]	Nucleic acid (pairs) type	MB, DB, NIN
		Nucleotide (pairs) frequency	LPF
		Evolutionarily derived features	KNN-derived features
	iDNA6mA [26]	Nucleic acid (pairs) type	MB
	SNNRice6mA [27]	Nucleic acid (pairs) type	MB
2020	p6mA [28]	Nucleotide (pairs) frequency	PSTNPds
		Physicochemical property	EIIP, PseKNC
	6mA-RicePred [29]	Nucleic acid (pairs) type	MB
		Nucleotide (pairs) frequency	Kmer, Markov features
		Physicochemical property	NCP
	6mA-Finder [30]	Nucleic acid (pairs) type	MB
		Nucleotide (pairs) frequency	ANF, CKSNAP, DNC, ENAC, NAC, TNC
		Physicochemical property	NCP, PseDNC, EIIP
	6mAPred-FO [31]	Nucleic acid (pairs) type	MB
		Nucleotide (pairs) frequency	Kmer, PseDNC
	Meta-i6mA [32]	Nucleic acid (pairs) type	MB, DB
		Nucleotide (pairs) frequency	NAC, DNC, TNC, Kmer, KNC
		Physicochemical property	NCP, EIIP, DPP
	DNA6mA-MINT [33]	Nucleic acid (pairs) type	MB
2021	iRicem6A-CNN [34]	Nucleic acid (pairs) type	DB
	i6mA-VC [35]	Nucleic acid (pairs) type	MB
		Nucleotide (pairs) frequency	Kmer
		Physicochemical property	NCP
	6mA-Pred [36]	Natural language processing	Kmer word segmentation
	Deep6mA [37]	Nucleic acid (pairs) type	MB
	i6mA-CNN [38]	Nucleic acid (pairs) type	MB, DB
		Nucleotide (pairs) frequency	ANF
		Physicochemical property	NCP, DPP
	SMEP [39]	Nucleic acid (pairs) type	MB
	6mAPred-MSFF [40]	Nucleic acid (pairs) type	1-gram, 2-gram
		Nucleotide (pairs) frequency	NAC, DNC

Note: NCP, nucleotide chemical property. MB, monomer binary. PseKNC, pseudo k-tuple nucleotide composition. KSNPF, k-spaced nucleotide pairs frequency. DPP, dimer physicochemical property. DB, dimer binary. NIN, numerical information of nucleotides. LPF, local position-specific dinucleotide frequency. KNN-derived features, k-nearest neighbors-derived features. PSTNPds, position-specific trinucleotide propensity based on double-strand. EIIP, electron-ion interaction potential. ANF, accumulated nucleotide frequency. CKSNAP, composition of k-spaced nucleic acid pairs. DNC, dinucleotide composition. ENAC, enhanced nucleic acid composition. NAC, nucleic acid composition. TNC, trinucleotide composition. PseDNC, pseudo dinucleotide composition. KNC, k-tuple nucleotide composition.

Independent datasets

After training and tuning of a model, independent dataset is often used to check the generalization ability of the model. Two types of independent datasets are prepared in this work:

- Same species independent dataset. Wang *et al.* [39] collected, processed and obtained a 6mA dataset in rice from eRice database [50], with positive and negative samples approaching 600 000, respectively. Considering the huge amount of samples, only 10 000 positive and negative samples are randomly selected, respectively, to form independent dataset of the same species. The samples of independent dataset have no intersection with the samples of the benchmark datasets Rice:Chen and Rice:Lv.
- Cross species independent datasets. Three cross species from 6mAPred-MSFF [40], including *A.thaliana*, *D.melanogaster* and *R.chinensis*, are directly adopted as independent datasets.

The details of the above datasets are shown in Table 2. All the sequences are 41 bp. The 21*th* position of the positive and negative sample sequence represents 6mA site and non-6mA site (the N6 position of adenine is not modified by -CH_3), respectively.

Architecture of MGF6mARice

section 1 shows the flow diagram of MGF6mARice's architecture. It mainly contains four modules: SMILES representation of DNA, MGF encoding, extracting features by residual blocks and MGF6mARice prediction.

Table 2. Details of all datasets

Type	Dataset	Number of Positives	Number of Negatives	Total number
Benchmark datasets	Rice:Chen	880	880	1,760
	Rice:Lv	154,000	154,000	308,000
Imbalanced datasets	Rice:Chen 1:5	176	880	1,056
	Rice:Chen 1:10	88	880	968
	Rice:Chen 1:20	44	880	924
	Rice:Lv 1:5	30,800	154,000	184,800
	Rice:Lv 1:10	15,400	154,000	169,400
	Rice:Lv 1:20	7,700	154,000	161,700
Independent datasets	NIP_10000	10,000	10,000	20,000
	A.thaliana	15,937	15,937	31,874
	D.melanogaster	11,191	11,191	22,382
	R.chinensis	11,815	11,815	23,630

Note: 1:5, 1:10, and 1:20 represent the random selection ratio from positive samples of Rice:Chen and Rice:Lv, respectively.

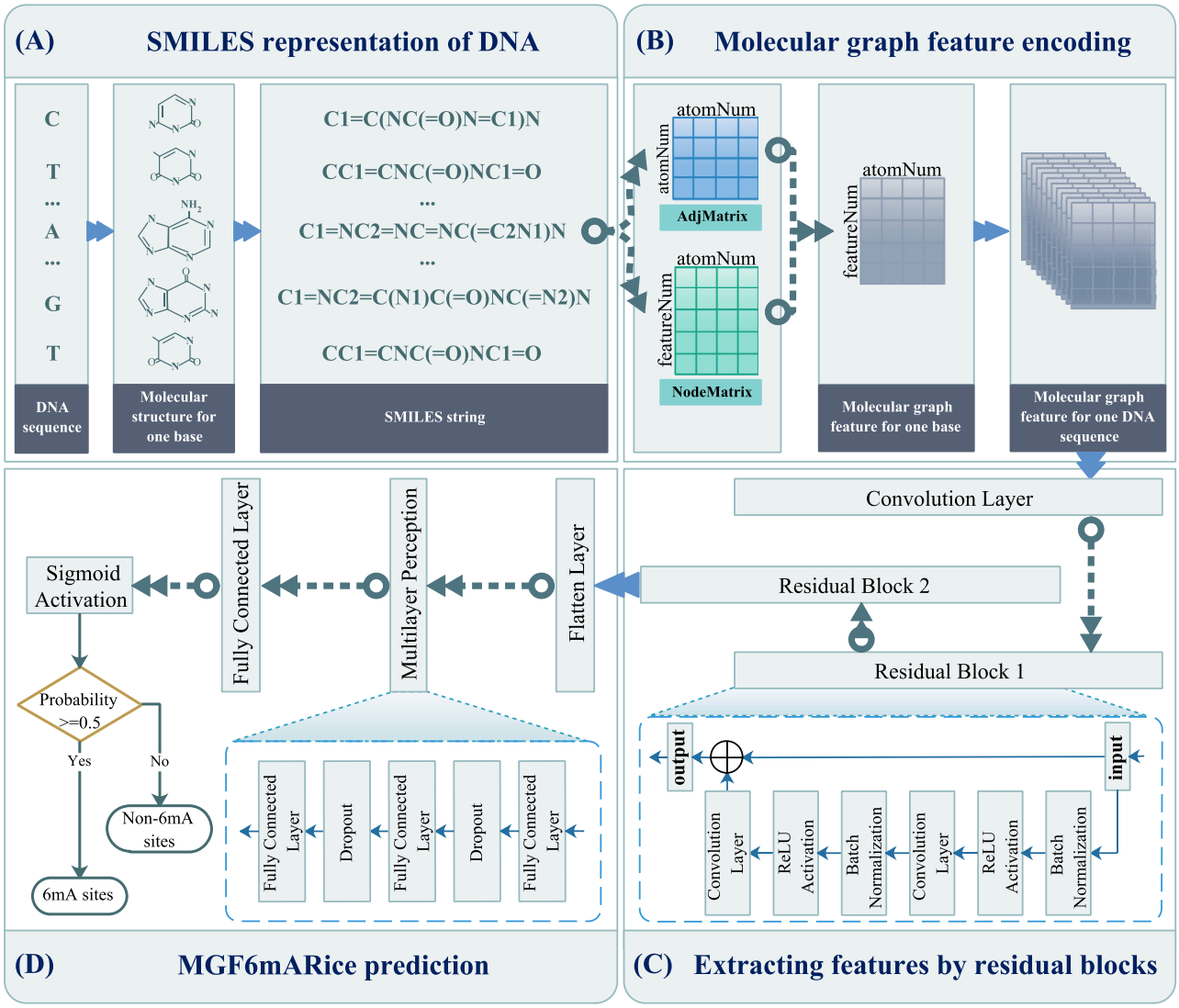

Figure 1. A flowchart of MGF6mARice. **(A)** SMILES representation of DNA. **(B)** Molecular graph feature encoding. AdjMatrix denotes adjacency matrix, NodeMatrix is node feature matrix. The shape of AdjMatrix is (atomNum, atomNum), where atomNum represents the number of atoms in the base molecule. The shape of NodeMatrix is (atomNum, featureNum), where featureNum is the dimension of the atomic features. Then multiplication of these two matrices is used to obtain the molecular graph feature of one base. By appending together molecular graph features of bases in order, the molecular graph feature is obtained for one DNA sequence. **(C)** Extracting features by residual blocks. Residual Block 1 and 2 denote the first and the second residual block, respectively. **(D)** MGF6mARice prediction.

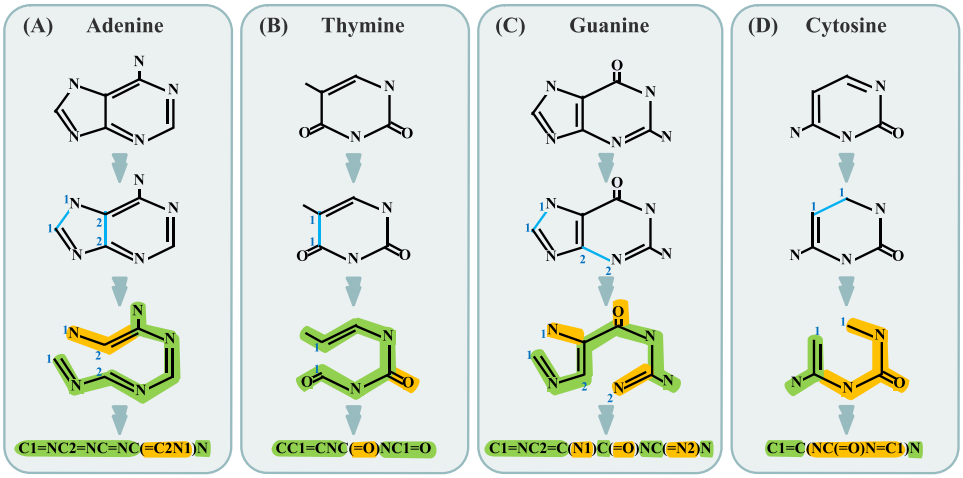

Figure 2. The SMILES representation of DNA bases. The blue bonds are assumed to be broken. The numbers denote the order of breakpoints in a molecule. The green and yellow highlights indicate the main and branch chain (in parentheses) formed after broken, respectively.

SMILES representation of DNA

In our work, how to effectively represent the molecular structure is the key step to realizing the feature encoding. Built by using molecular graph, SMILES is a concise chemical molecular language to describe molecular structure via ASCII string [46]. SMILES string has been widely used in bioinformatics and chemoinformatics, such as interaction and binding affinity prediction of drug–target [51–53], peptide toxicity prediction [43] and so on. For a DNA sequence, there are four kinds of bases, i.e. adenine (A), thymine (T), guanine (G) and cytosine (C). Each base corresponds to a chemical molecular structure. The chemical molecular structure needs to be converted into a form that can be processed by computer. Nevertheless, the SMILES string has not been applied on DNA bases to further extract features so far. Therefore, we utilize the SMILES string in this section to represent DNA, as shown in Figure 1A. The specific process of converting the chemical molecular structure of bases to the corresponding SMILES string is shown in Figure 2. Note that the canonical SMILES [54] is used in this study to avoid ambiguity.

Take base A as an example (Figure 2A): (i) The first row shows the chemical molecular structure of base A, including carbon atom (C, expressed by inflection point or endpoint), nitrogen atom (N), single bond and double bond (=). (ii) In the molecular structure on the second row, two blue bonds are assumed to be broken and denoted by 1, 2 respectively. Thus, a closed molecular structure is transformed into a linear open form for an easier description. (iii) In the third row, there are the main chain (marked in green) and the branch chain (marked in yellow). (iv) Finally, all atoms, bonds and breakpoints are recorded in the prescribed order, where these at the branch chain are enclosed in parentheses. So the SMILES representation of base A is obtained as: C1=NC2=NC=NC(=C2N1)N. SMILES strings of other bases are obtained in the same way (Figure 2B–D).

Therefore, we obtain the related data from PubChem database [55] to construct the set of SMILES strings to characterize the DNA sequence.

MGF encoding

Given SMILES strings, it is a crucial step to encode features with biological significance. In the graph structure data, the neighbors of nodes are very important for the representation of nodes [56, 57]. In this study, we integrate the neighbors of atoms (i.e. nodes) to encode features for one base in the DNA sequence. In order to extract effective information from the neighbors of atoms, we construct the MGF encoding in terms of the following three steps.

First, we construct a graph of DNA sequence $G_s = \{G_1, G_2, \cdots, G_l\}$, where l is the number of bases in this DNA sequence. $G_b = (V_b, E_b)$ represents the corresponding graph of base b, where $V_b = \{v_1, v_2, \cdots, v_m\}$ is the set of all atoms and m represents the number of atoms, and $E_b = \{(v_1, v_2), (v_1, v_3), \cdots, (v_i, v_j)\}$ is the set of connections between atoms.

Then, we construct an adjacency matrix $\mathbf{A}$ and a node feature matrix $\mathbf{N}$ of base, respectively. The element a_{ij} of $\mathbf{A}$ denotes whether there is a connection between atoms or not. If (v_i, v_j) belongs to E_b, a_{ij} is 1, otherwise a_{ij} is 0. There is no connection between the atom itself, that is, the main diagonal is 0. Since the dimension of $\mathbf{A}$ is only determined by m in G_b, $\mathbf{A} \in \mathbb{R}^{m \times m}$, $\mathbf{A}$ is also a symmetric matrix.

However, the adjacency matrix $\mathbf{A}$ can only reflect 2D connection information between atoms. The atomic features also embody the 3D structure information of the molecule [58], which can be used to increase interpretability. Hence, the features of atoms are adopted in the form of the node feature matrix $\mathbf{N}$. One row vector $\boldsymbol{n_i}$ of $\mathbf{N}$ represents all the features of an atom v_i in G_b. The length of $\boldsymbol{n_i}$ equals the dimension of the atomic features, k. Therefore, the shape of $\mathbf{N}$ is $\mathbb{R}^{m \times k}$. The atomic features

used in this work include the atomic symbol, the degree of the atom in the molecule and so on [59]. The details can be seen in Supplementary Table S1.

GCN is a promotion of convolution on graph structure data, which can fully integrate node information, and has been widely adopted by bioinformatics [42, 57, 60]. Its essence is to extract the structural features of the graph [47]. In this section, for the adjacency matrix $\mathbf{A}$ and node feature matrix $\mathbf{N}$, we extract the DNA MGF by using GCN inherent principles. To be specific, for the MGF of one base $\boldsymbol{MGF}_{base}$, it can be calculated by $\boldsymbol{MGF}_{base} = \hat{\mathbf{A}}\hat{\mathbf{N}}$, where $\hat{\mathbf{A}}$ and $\hat{\mathbf{N}}$ are the normalized matrix, i.e.

$$\hat{\mathbf{A}} = \boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{A}+\boldsymbol{I})\boldsymbol{D}^{-\frac{1}{2}}, \tag{1}$$

$$\hat{\boldsymbol{n}}_i = \frac{\boldsymbol{n}_i}{\sum_{j=0}^{k} n_{ij}}, \tag{2}$$

where $\boldsymbol{I}$ is the identity matrix with the same size of $\mathbf{A}$, which is used to add self-loop to include the feature of the node itself (i.e. atom in molecule). $\boldsymbol{D}$ is the degree matrix of $\mathbf{A}+\boldsymbol{I}$. n_{ij} is an element in the node feature matrix $\mathbf{N}$. $\hat{\boldsymbol{n}}_i$ represents one row vector of the normalized node feature matrix $\hat{\mathbf{N}}$.

Finally, the MGF of one DNA sequence $\boldsymbol{MGF}_{seq}$ is obtained by appending together the MGFs of bases in order (Figure 1B). The details are provided in Supplementary Materials.

Extracting features by residual blocks

Many researches have shown that, compared with the direct connection between layers in the traditional neural network, the residual block by shortcut connection in ResNet can effectively prevent the gradient from exploding or disappearing in deep network [61, 62]. Inspired by He *et al.*, we use the combination of residual blocks to mine the useful information for 6mA prediction in MGF (Figure 1C). First, we input the $\boldsymbol{MGF}_{seq}$ into a convolutional layer (*Conv*) for preliminary feature extraction, and then feed the output into two residual blocks to extract more effective and distinguishable features:

$$O_{conv} = Conv(\boldsymbol{MGF}_{seq}) \tag{3}$$

$$O_{r1} = O_{conv} + Conv_{relu}(BN(Conv_{relu}(BN(O_{conv})))) \tag{4}$$

$$O_{r2} = Conv(O_{r1}) + Conv_{relu}(BN(Conv_{relu}(BN(MP(O_{r1}))))) \tag{5}$$

where O_{conv}, O_{r1} and O_{r2} are the output of the convolutional layer, the first and the second residual block, respectively. $Conv_{relu}$ is a convolution layer with *relu* activation function. *BN* denotes batch normalization. *MP* denotes a max-pooling layer.

MGF6mARice prediction

For the deeper features extracted using residual blocks, we use multilayer perception (MLP) to build a prediction module to determine whether the DNA sequence contains 6mA site (Figure 1D). The output of residual blocks (O_{r2}) is fed into three-layer MLP with *relu* activation function (FC_{relu}), i.e.

$$\boldsymbol{y} = FC_{relu}(FC_{relu}(FC_{relu}(O_{r2}))) \tag{6}$$

Finally, the probability of a sample being a 6mA site (*pred*) is calculated as follows:

$$pred = FC_{sigmoid}(\boldsymbol{y})), \tag{7}$$

where $FC_{sigmoid}$ denotes a fully connected layer with a *sigmoid* activation function. If *pred* is less than 0.5, the DNA sequence is a negative sample, which means that there is no 6mA site. Otherwise, it is a positive sample that contains a 6mA site.

Optimization of MGF6mARice

The optimization of hyperparameters plays a vital role in the prediction model of neural networks [42, 63]. In our predictive model, the hyperparameters to be optimized are the number of residual blocks, number of fully connected layers in MLP, number of units in each fully connected layer, dropout rate, optimizer and batch size (see Supplementary Table S2 for value range). To reduce optimization time and improve optimization efficiency, we use the Bayesian optimization algorithm to obtain optimized hyperparameters, provided by the Python package hyperopt [64]. The details of the relevant results are given in Supplementary Table S3.

In addition, we utilize the binary cross-entropy as the loss function and use the stochastic gradient descent optimizer to optimize it. To prevent overfitting, the batch normalization and dropout layer are adopted in MGF6mARice. At the same time, MGF6mARice also uses a learning rate decay strategy to facilitate the optimization and generalization of the model [65].

Evaluation metrics

In this study, 10-fold cross-validation (10-CV) is adopted to evaluate the performance of MGF6mARice and state-of-the-art methods. All samples are randomly divided into the training set, validation set and test set in a ratio of 8:1:1 in both the benchmark and imbalanced datasets. The average of 10 test set results is used as the final 10-CV result. To assess the performance of our proposed method and others, several traditional evaluation measures [66, 67] are employed, including sensitivity (Sn), specificity (Sp), accuracy (Acc) and Mathew's correlation coefficient (MCC). These metrics can be formulated as

Table 3. A comparison of 10-fold cross-validation performance for the various methods

Dataset	Methods	Sn	Sp	Acc	MCC	AUC
Rice:Chen	SNNRice6mA	0.93	0.96	0.94	0.89	0.98
	i6mA-CNN	0.87	0.87	0.87	0.74	0.93
	6mAPred-MSFF	0.95	0.95	0.95	0.90	0.98
	MGF6mARice	**0.96**	**0.97**	**0.97**	**0.93**	**0.99**
Rice:Lv	SNNRice6mA	0.94	0.89	0.91	0.83	0.97
	i6mA-CNN	0.95	0.93	0.94	0.88	0.98
	6mAPred-MSFF	**0.97**	0.93	0.95	0.90	0.99
	MGF6mARice	0.96	**0.96**	**0.96**	**0.92**	**0.99**

Note: for each method, the optimal result has been shown after running 20 and 10 times 10-fold cross-validation on Rice:Chen and Rice:Lv dataset, respectively. The best value of each index is bolded.

follows:

$$Sn = \frac{TP}{TP + FN} \tag{8}$$

$$Sp = \frac{TN}{TN + FP} \tag{9}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{11}$$

where TP (true positive) (TN: true negative) represents the number of correctly predicted positive (negative) samples. FP (false positive) (FN: false negative) denotes the number of incorrectly identified positive (negative) samples.

In addition, the AUC (area under the receiver-operating characteristic curve) and AUPR (area under the precision-recall curve) are commonly used to estimate the overall performance of a predictor. Among them, AUPR is the more reliable performance measure for imbalanced data [68].

Results and Discussion

Comparison with state-of-the-art methods

To evaluate the prediction performance of MGF6mARice, we compare it with the following state-of-the-art methods, including 6mAPred-MSFF [40], i6mA-CNN [38] and the classical method SNNRice6mA [27]. Afterward, on the aforementioned three types of datasets, multiple comparative experiments are carried out between MGF6mARice and comparison methods.

Comparison of performance on benchmark datasets

As the size of Rice:Chen dataset is smaller than that of Rice:Lv dataset, 10-CV is performed 20 times on Rice:Chen and 10 times on Rice:Lv. Table 3 tabulates the results for four methods.

It can be seen that MGF6mARice is superior to other methods. Specifically, compared with the second-best result, 6mAPred-MSFF, for the performance indices Sp, Acc and MCC, MGF6mARice still achieves 2%, 2% and 3% improvement, respectively, on Rice:Chen; and achieves 3%, 1% and 2% improvement, respectively, on Rice:Lv.

Simultaneously, in order to investigate the robustness of the various methods, the boxplot is drawn for observation. Figure 3 shows the dispersion of the 10-CV results (20 times) on Rice:Chen.

We can see that MGF6mARice is more robust than other methods. The visualization of the dispersion of the 10-CV results (10 times) on Rice:Lv is shown in Figure 4.

It can be seen that the robustness of MGF6mARice is slightly decreased. Nevertheless, combining the results on two benchmark datasets, the robustness of MGF6mARice has little variation compared with other methods. This indicates that MGF6mARice is relatively less affected by the size of dataset and more stable.

In conclusion, the performance of MGF6mARice is better than that of other methods considering accuracy and robustness for two benchmark datasets.

Comparison of performance on imbalanced datasets

Figure 5 displays the AUPR of each comparative method upon 10-CV. It can be seen that, as the gap between the number of positive and negative samples grows larger, AUPR decreases significantly, because of sensitivity to the imbalance ratio. Under different ratios of positive and negative samples, whether on a large dataset (Rice:Lv) or a small dataset (Rice:Chen), the AUPR of MGF6mARice is better than other methods.

Comparison of performance on independent datasets

The experimental results on independent datasets are used to further judge the generalization performance of MGF6mARice and other methods. We can see from Table 4 that the accuracy of MGF6mARice is better than that of other methods. Therefore, MGF6mARice has a relatively good generalization performance compared with other methods.

Effectiveness of MGF

The proposed method MGF6mARice exploits a newly constructed feature MGF, which is extracted from the

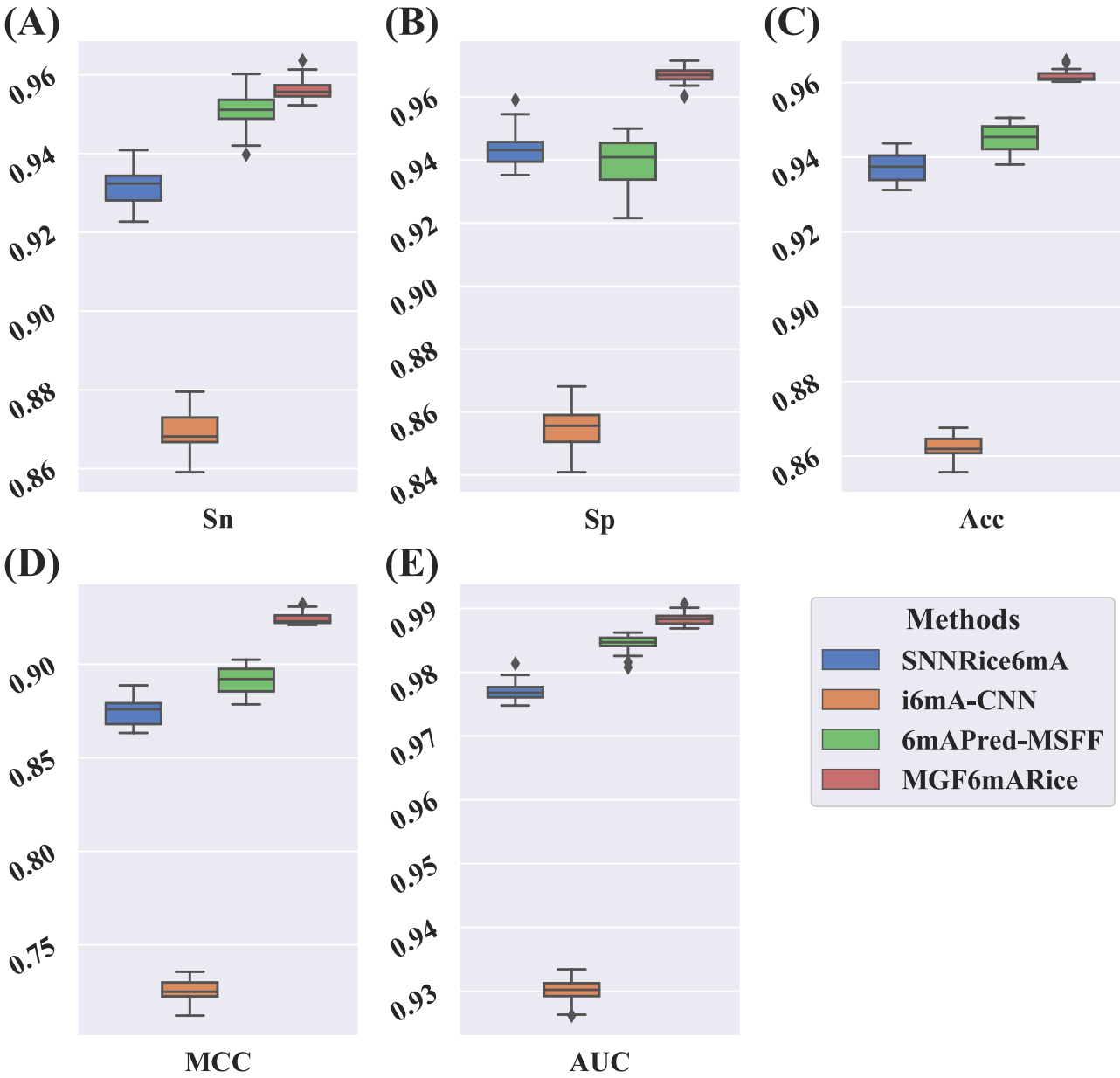

Figure 3. The dispersion of MGF6mARice and the compared methods under 20 times 10-fold cross-validation results on Rice:Chen. **(A)–(E)** indicate the average of Sn, Sp, Acc, MCC and AUC, respectively.

chemical structure of base from the DNA sequence, to predict the binary classification problem of 6mA sites in rice. To evaluate the validity of the MGF, eight traditional and typical features are utilized for comparison experiments, which were derived from previous studies. These compared features can be divided into three categories: (a) nucleic acid type-based, including MB, dimer binary (DB), 1-gram and 2-gram encoding; (b) frequency of nucleic acid type-based, including nucleic acid composition (NAC) and dinucleotide composition (DNC); (c) physicochemical property-based, such as NCP and dimer physicochemical property (DPP). These features are obtained by referring to iLearn [69].

Firstly, the network of MGF6mARice, denoted as MGF6mARice-Model, is used as the classifier for the various features. For a fair comparison, these comparative experiments are carried out under the same training set and test set, by using 10-CV (only perform on the Rice:Chen dataset because of the time and space cost on the Rice:Lv dataset). Figure 6A shows the comparison results of five metrics (Sn, Sp, Acc, MCC and AUC).

Moreover, we also use five common classifiers, i.e. k-nearest neighbors (KNN), SVM, MLP, logistic regression (LR) and naive bayes (NB), to construct prediction models for various features. Figure 6B–F show the experimental results. We can see from Figure 6 that, for whichever classifier, the performance of the MGF is significantly superior to other features. Therefore, the MGF is more suitable for the 6mA prediction problem than other features. The possible reason is that DNA N6-methyladenine modification occurs at the molecular level. The extracted feature from the chemical molecular structure of the

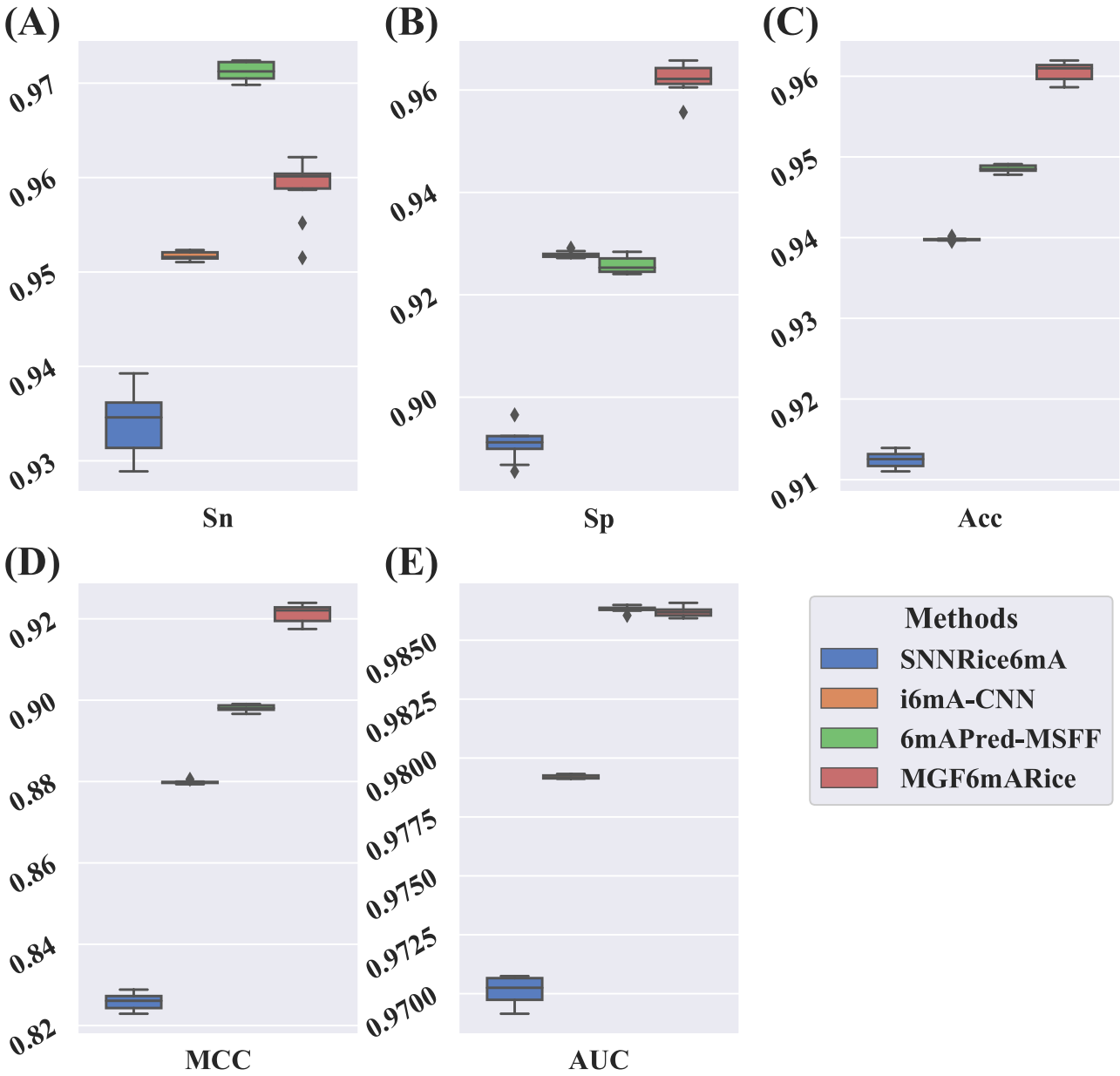

Figure 4. The dispersion of MGF6mARice and the compared methods under 10 times 10-fold cross-validation results on Rice:Lv. **(A)–(E)** indicate the average of Sn, Sp, Acc, MCC and AUC, respectively.

DNA base is more close to the mechanism of chemical molecular generation, and more biologically meaningful than other traditional features. Thus, it is feasible and meaningful to explore effective molecular representation from chemical microstructure.

Ablation study

The effectiveness of MGF6mARice has been verified in the above sections. To further illustrate the contribution of the remaining main parts of MGF6mARice, we conduct a series of ablation studies by considering different modular combinations. Considering the huge amount of data in Rice:Lv dataset and the high time cost, the ablation studies are only performed on the Rice:Chen dataset. The following variants of MGF6mARice are taken into account:

- MGF6mARice-MLP is the variant without MLP from the output module.
- MGF6mARice-Res is the variant without residual blocks.
- MGF6mARice-Res-MLP is the variant without MLP and residual blocks.

Table 5 tabulates the performance of MGF6mARice and three variants under five metrics computing with 10-CV.

We can observe that the performance indices are all decreased when removing any component of MGF6mARice. That is to say, these components are essential to MGF6mARice. To be specific, when without residual blocks and MLP, MGF6mARice-Res-MLP has the greatest performance degradation compared with MGF6mARice (the Acc, MCC and AUC decreased by 9%, 16% and 4%,

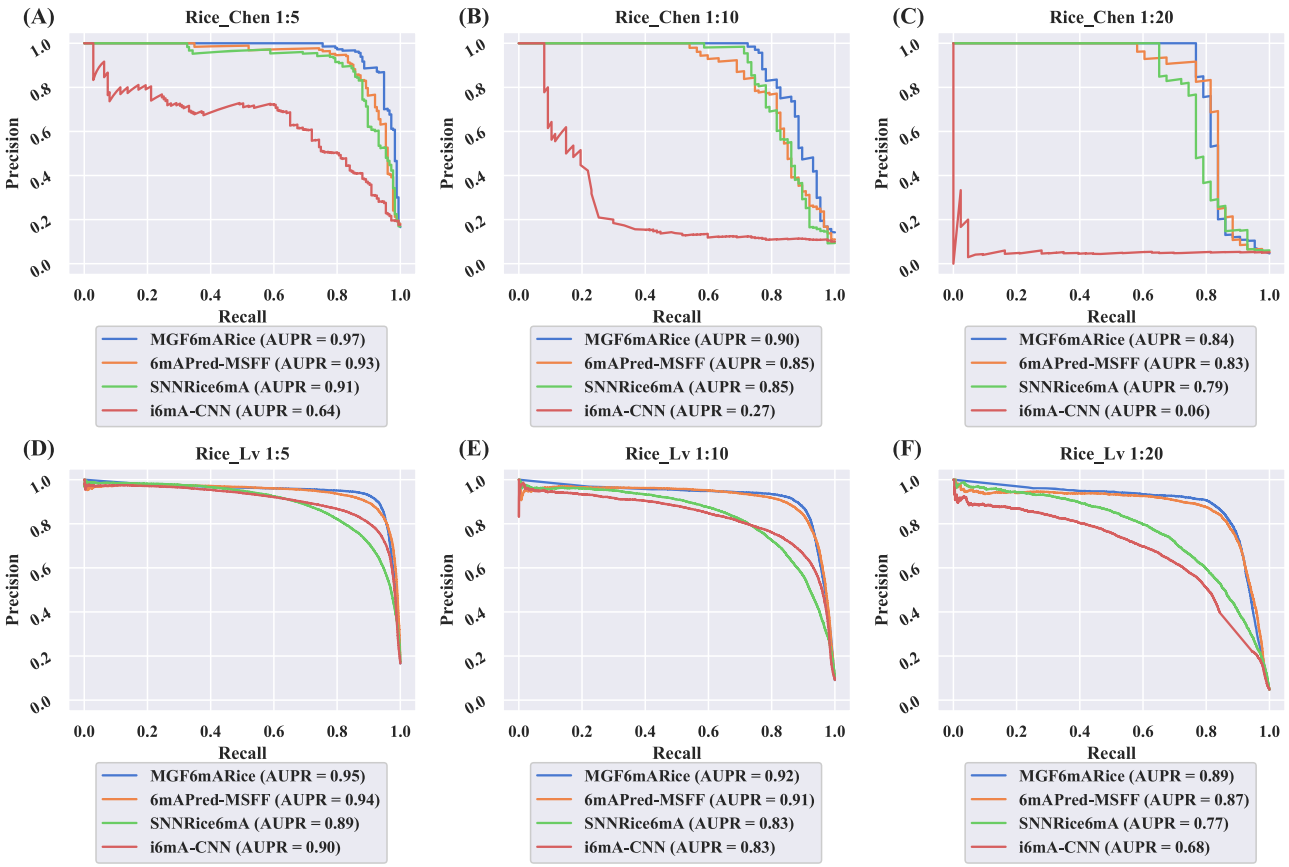

Figure 5. The Precision-Recall curves of MGF6mARice and other methods on six imbalanced datasets **((A)–(F))** using 10-fold cross-validation. 1:5, 1:10 and 1:20 represent the random selection ratio from positive samples of Rice:Chen and Rice:Lv, respectively.

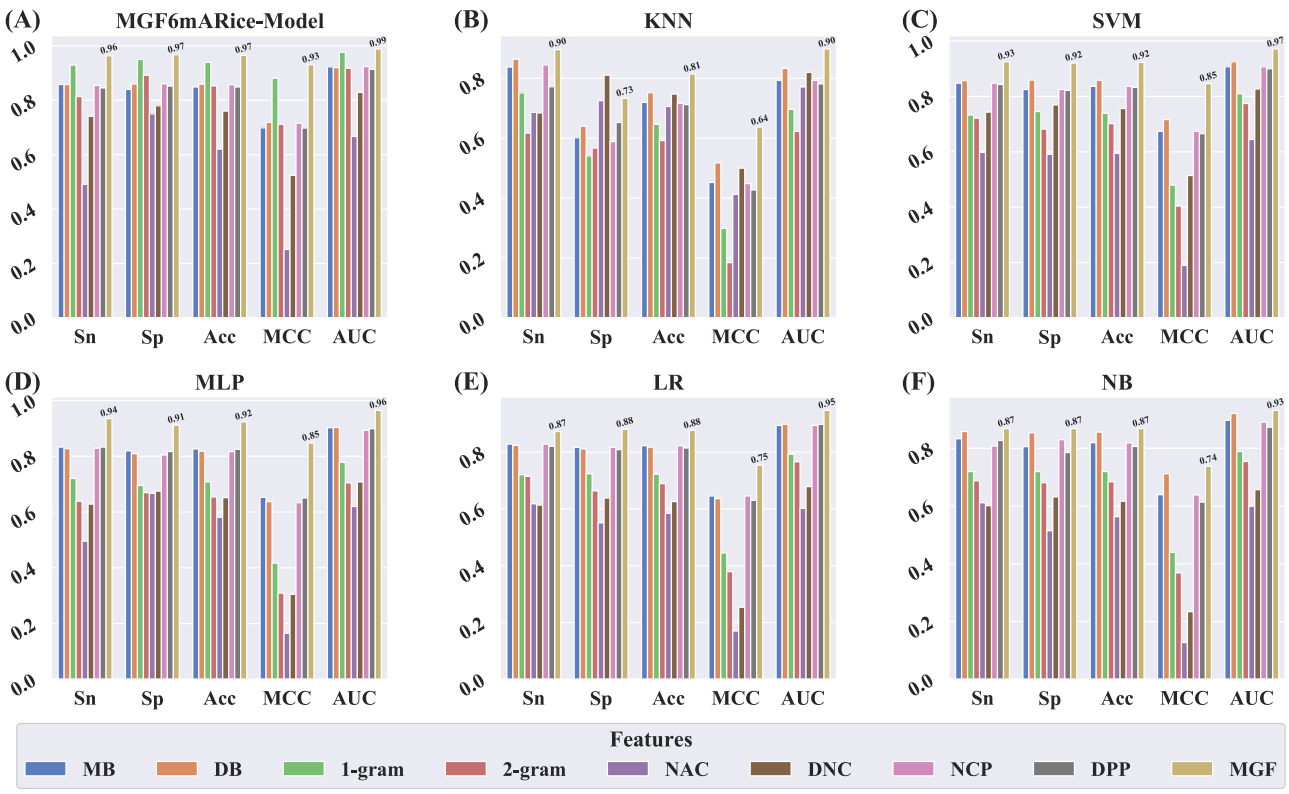

Figure 6. The 10-fold cross-validation results of the various features (MB, DB, 1-gram, 2-gram, NAC, DNC, NCP, DPP and MGF (this study)) on six classifiers. **(A)** Results of MGF6mARice-Model (the network of MGF6mARice); **(B)** Results of KNN; **(C)** Results of SVM; **(D)** Results of MLP; **(E)** Results of LR; **(F)** Results of NB.

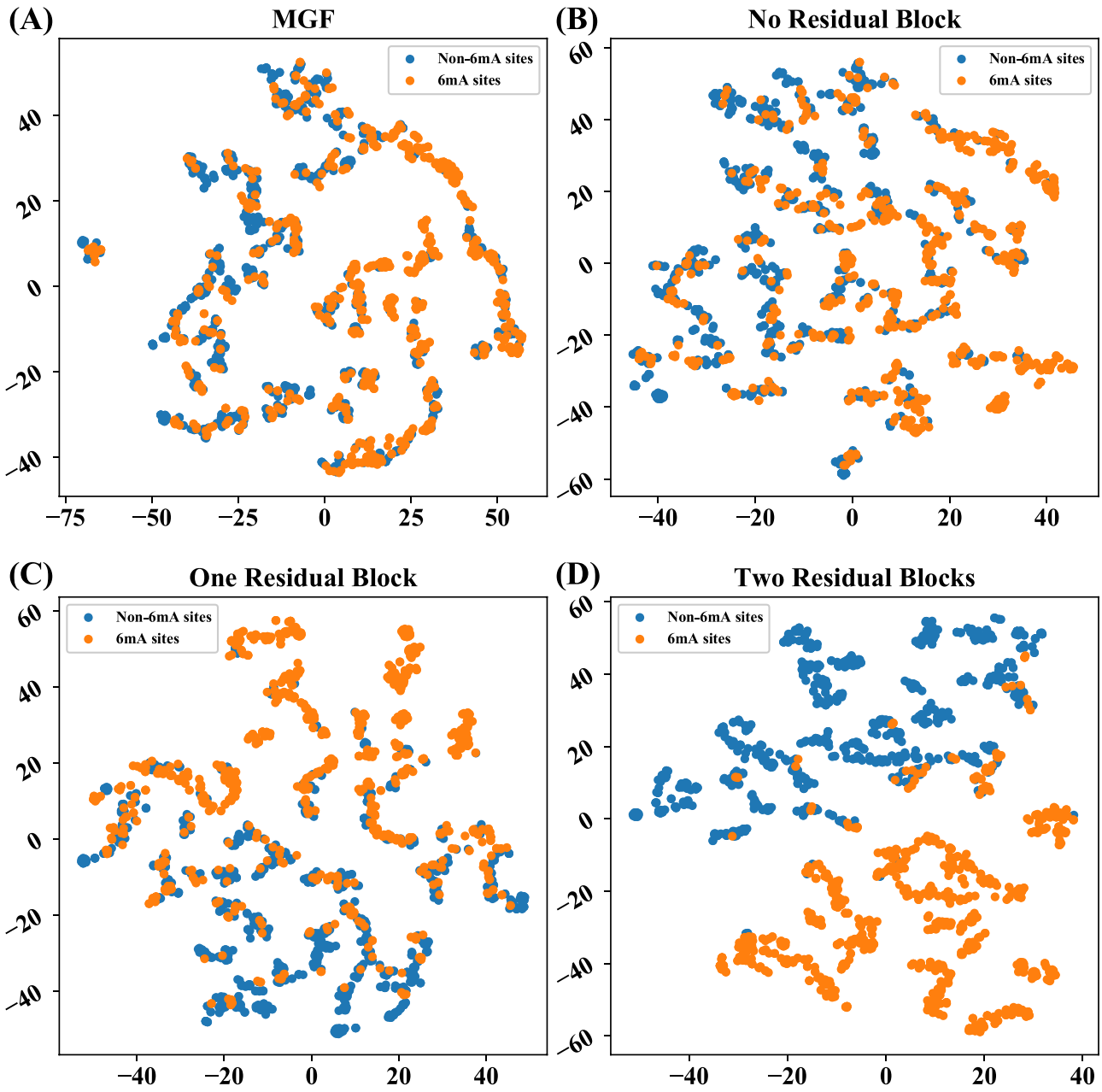

Figure 7. The t-SNE visualization of features extracted by residual blocks. **(A)** Visualization of MGF; **(B)** Visualization of features extracted without residual block; **(C)** Visualization of features extracted by one residual block; **(D)** Visualization of features extracted by two residual blocks. Orange and blue color indicates 6mA and non-6mA samples, respectively.

respectively). After using MLP, i.e. MGF6mARice-Res, compared with MGF6mARice-Res-MLP, the Acc, MCC and AUC are increased by 4%, 7% and 1%, respectively.

For the comparison between MGF6mARice-MLP and MGF6mARice-Res-MLP, in other words, utilizing residual blocks, we can see that the performance has been greatly enhanced (the Acc and MCC raised by 7% and 14%, respectively, the AUC achieves 0.99). Through the ablation experiments, it can be seen that residual blocks contribute the most to MGF6mARice.

To observe the effectiveness of the residual block more intuitively, t-SNE (t-distributed stochastic neighbor embedding) [70] is applied to visualize their distribution via reducing the features extracted by the residual block to two-dimension (2D) (presented in Figure 7). Figure 7A–B show that, based on MGF, the preliminary features extracted by a convolutional layer, i.e. without residual block, do not yet have good classification performance. Compared with no residual block, the features extracted from one residual block (Figure 7C) have better classification performance. When the number of residual blocks increases to two (Figure 7D), the extracted features can make a good division of 6mA samples and non-6mA samples. It indicates that two residual blocks already have a strong discriminating performance for the DNA 6mA sites classification problem in rice.

In addition, visualization about the classification of MGF6mARice is performed to intuitively reflect

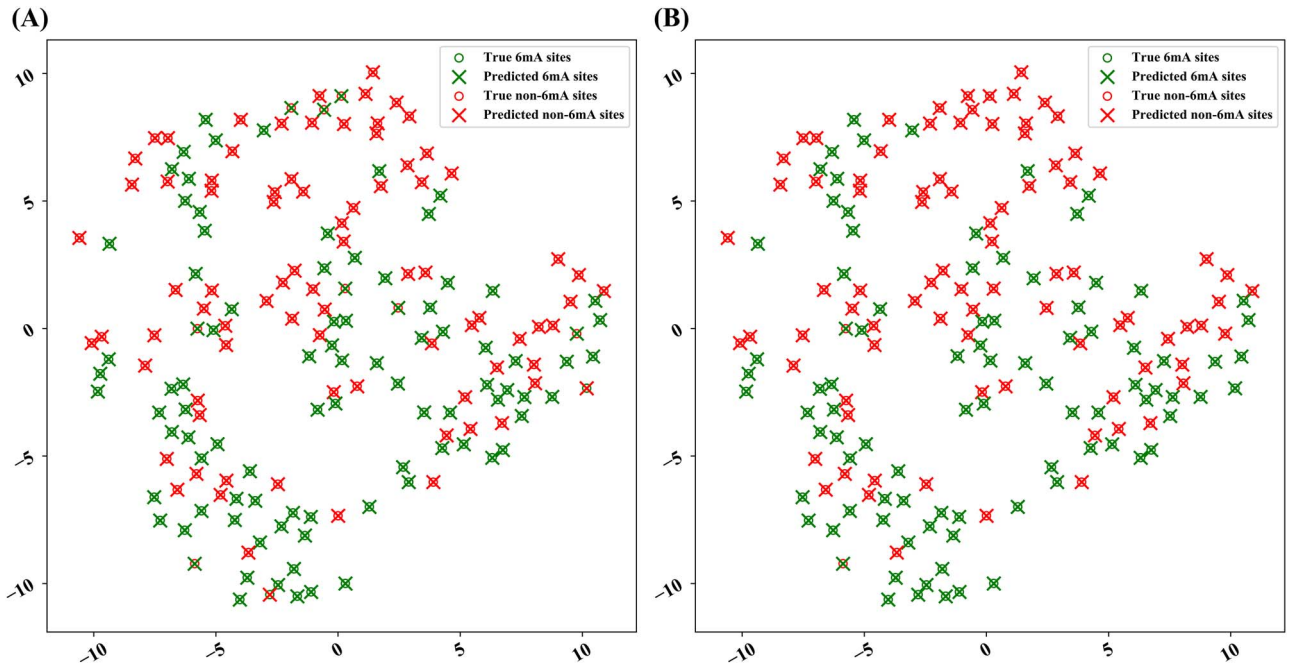

Figure 8. The t-SNE visualization of the predicted sample category from the MGF6mARice classification and the real sample category. Green and red colors indicate the data distribution of the original positive and negative samples, respectively. The circles represent true samples categories and the crosses represent prediction categories. When the predicted category is consistent with the real category, that is, when the circle and cross colors are the same, the prediction is correct, otherwise the prediction is wrong. **(A)** represents the visualization of the classification by MGF6mARice without residual blocks. **(B)** indicates the visualization of the classification by MGF6mARice. Note that considering that a large amount of data will lead to severe occlusion, one-tenth of the samples in the Rice:Chen dataset are randomly selected for visual display [i.e. the number of samples on (A) and (B) is both 176, of which positive and negative samples is both 88].

Table 4. The prediction accuracy of the various methods on four independent datasets

Dataset	Methods	Acc
NIP_10000	SNNRice6mA	0.81
	i6mA-CNN	0.87
	6mAPred-MSFF	0.82
	MGF6mARice	**0.88**
A.thaliana	SNNRice6mA	0.70
	i6mA-CNN	0.75
	6mAPred-MSFF	0.73
	MGF6mARice	**0.77**
D.melanogaster	SNNRice6mA	0.75
	i6mA-CNN	0.81
	6mAPred-MSFF	0.79
	MGF6mARice	**0.82**
R.chinensis	SNNRice6mA	0.79
	i6mA-CNN	0.85
	6mAPred-MSFF	0.83
	MGF6mARice	**0.86**

Note: the optimal value is bolded.

Table 5. The performance of MGF6mARice and its variants using 10-fold cross-validation

	Sn	Sp	Acc	MCC	AUC
MGF6mARice	**0.96**	**0.97**	**0.97**	**0.93**	**0.99**
MGF6mARice-MLP	0.95	0.96	0.95	0.91	0.99
MGF6mARice-Res	0.90	0.94	0.92	0.84	0.96
MGF6mARice-Res-MLP	0.88	0.88	0.88	0.77	0.95

Note: the optimal value of each metric is bolded.

the contribution of the residual block on the final classification results. We use t-SNE to visualize the original data distribution and then compare the real sample category with the predicted sample category from the MGF6mARice classification. The results are displayed in Figure 8. Figure 8A represents the visualization of classification by MGF6mARice without residual blocks and Figure 8B indicates the visualization of classification by MGF6mARice. Comparing Figure 8A and Figure 8B, it can be found that the number of misclassifications by MGF6mARice is significantly reduced after using the residual block. The effectiveness of the residual block is again demonstrated based on the above visualization of classification results.

Conclusions

DNA methylation is one of the common epigenetic modifications by which methyl groups are added to the molecule of DNA base, without changing the DNA sequence. Among them, 6mA in rice is involved in many biological processes, such as inhibiting transcriptional regulation of gene expression, regulating rice growth and development and stress response. 6mA sites prediction in rice will help accelerate and assist researchers in relevant biological analysis. In the current research, using DNA MGF and residual block, we propose a 6mA prediction method for rice called MGF6mARice. Inspired by the generation mechanism of 6mA, utilizing the principle of GCN, MGFs are mined and calculated from the

chemical structure of DNA bases. For MGFs obtained, to further extract higher level, easy-to-recognize features, the residual block structure is considered. Compared with state-of-the-art methods, experimental results on various datasets demonstrate that MGF6mARice achieves better comprehensive performance. Moreover, comparing different feature encodings on the six classifiers illustrates the suitability and effectiveness of the MGF in the 6mA sites prediction problem of rice. Furthermore, through the ablation study and t-SNE visualization of the residual block, it is proved that the residual block can greatly increase the performance of MGF6mARice. The results of various experiments have certificated that MGF6mARice is a tool full of promise for distinguishing 6mA sites in rice.

Although MGF6mARice has achieved well-predictive performance, there are still some improvements. For instance, i6mA-CNN not only developed a tool to predict 6mA sites in rice but also provided a potential motif identification algorithm. Therefore, in the future, MGF6mARice can also consider adding the motif recognition function of 6mA. In this study, only using a single feature, the combination of the MGF constructed here with other features can also be considered in the future. Additionally, MGF6mARice, as a tool for predicting 6mA sites, utilizing DNA MGF, can also be transferred to other methylation type predictions, in terms of the nucleic acid sequence.

Key Points

- A new deep learning method is proposed based on the DNA molecular graph feature and residual block, which improves the prediction performance of DNA 6mA sites in the rice genome.
- A series of comparison experiments are performed by using different features and classifiers. In terms of the prediction performance, the DNA molecular graph feature, calculated by the GCN principle from the chemical molecular structure of DNA bases, is more suitable and effective for DNA 6mA site prediction.
- The residual block is used to make the features rich, and obtain higher level and easy-to-distinct features. Ablation study and t-SNE visualization show that it can greatly improve prediction performance.
- To the best of our knowledge, it is the first time to devise a feature of DNA sequence by considering the chemical molecular structure.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgments

We thank anonymous reviewers for valuable suggestions.

Data availability

The data set and source code can be freely downloaded from https://github.com/zhlSunLab/MGF6mARice.

Funding

This work was supported by a National Natural Science Foundation of China (No. 61972002).

References

1. Heard E, Martienssen RA. Transgenerational epigenetic inheritance: myths and mechanisms. *Cell* 2014;**157**(1):95–109.
2. Lv H, Dao FY, Zhang D, *et al.* Advances in mapping the epigenetic modifications of 5-methylcytosine (5mC), N6-methyladenine (6mA), and N4-methylcytosine (4mC). *Biotechnol Bioeng* 2021;**118**(11):4204–16.
3. Ye P, Luan Y, Chen K, *et al.* MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. *Nucleic Acids Res* 2016;**45**(D1):gkw950.
4. O'Brown ZK, Greer EL. N6-methyladenine: a conserved and dynamic DNA mark. *Adv Exp Med Biol* 2016;**945**:213–46.
5. Du K, Zhang S, Chen W, *et al.* Epigenetic DNA modification N^6-methyladenine inhibits DNA replication by *Sulfolobus solfataricus* Y-family DNA polymerase Dpo4. *Arch Biochem Biophys* 2019;**675**:108120.
6. Li X, Zhu J, Hu F, *et al.* Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. *BMC Genomics* 2012;**13**(1):300.
7. Zhang Q, Liang Z, Cui X, *et al.* N^6-methyladenine DNA methylation in japonica and Indica rice genomes and its association with gene expression, plant development, and stress responses. *Mol Plant* 2018;**11**(12):1492–508.
8. Zhou S, Li X, Liu Q, *et al.* DNA demethylases remodel DNA methylation in rice gametes and zygote and are required for reproduction. *Mol Plant* 2021;**14**(9):1569–83.
9. Liang Z, Riaz A, Chachar S, *et al.* Epigenetic modifications of mRNA and DNA in plants. *Mol Plant* 2020;**13**(1):14–30.
10. Fu Y, Luo GZ, Chen K, *et al.* N^6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. *Cell* 2015;**161**(4):879–92.
11. Luo GZ, Blanco MA, Greer EL, *et al.* DNA N^6-methyladenine: a new epigenetic mark in eukaryotes? *Nat Rev Mol Cell Biol* 2015;**16**(12):705–10.
12. Laird PW. Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet* 2010;**11**(3):191–203.
13. Chachar S, Liu J, Zhang P, *et al.* Harnessing current knowledge of DNA N6-methyladenosine from model plants for non-model crops. *Front Genet* 2021;**12**:668317.
14. Zhu S, Beaulaurier J, Deikus G, *et al.* Mapping and characterizing N6-methyladenine in eukaryotic genomes using single-molecule real-time sequencing. *Genome Res* 2018;**28**(7):1067–78.
15. Zhou C, Wang C, Liu H, *et al.* Identification and analysis of adenine N^6-methylation sites in the rice genome. *Nat Plants* 2018;**4**(8):554–63.
16. Xiao CL, Zhu S, He M, *et al.* N^6-methyladenine DNA modification in the human genome. *Mol Cell* 2018;**71**(2):306–18.
17. Xu H, Jia P, Zhao Z. Deep4mC: systematic assessment and computational prediction for DNA N4-methylcytosine sites by deep learning. *Brief Bioinform* 2021;**22**(3):bbaa099.

18. Liu Q, Chen J, Wang Y, *et al.* DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites. *Brief Bioinform* 2021;**22**(3):bbaa124.
19. Chen W, Lv H, Nie F, *et al.* i6mA-Pred: identifying DNA N^6-methyladenine sites in the rice genome. *Bioinformatics* 2019;**35**(16):2796–800.
20. Le NQK. iN6-methylat (5-step): identifying DNA N^6-methyladenine sites in rice genome using continuous bag of nucleobases via Chou's 5-step rule. *Mol Genet Genomics* 2019;**294**(5):1173–82.
21. Pian C, Zhang G, Li F, *et al.* MM-6mAPred: identifying DNA N6-methyladenine sites based on Markov model. *Bioinformatics* 2020;**36**(2):388–92.
22. Lv H, Dao FY, Guan ZX, *et al.* iDNA6mA-Rice: a computational tool for detecting N6-methyladenine sites in rice. *Front Genet* 2019;**10**:793.
23. Liu Z, Dong W, Jiang W, *et al.* csDMA: an improved bioinformatics tool for identifying DNA 6 mA modifications via Chou's 5-step rule. *Sci Rep* 2019;**9**(1):13109.
24. Kong L, Zhang L. i6mA-DNCP: computational identification of DNA N^6-methyladenine sites in the rice genome using optimized dinucleotide-based features. *Gene* 2019;**10**(10):828.
25. Basith S, Manavalan B, Shin TH, *et al.* SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Mol Ther Nucleic Acids* 2019;**18**:131–41.
26. Tahir M, Tayara H, Chong KT. iDNA6mA (5-step rule): identification of DNA N6-methyladenine sites in the rice genome by intelligent computational model via Chou's 5-step rule. *Chemometr Intell Lab* 2019;**189**:96–101.
27. Yu H, Dai Z. SNNRice6mA: a deep learning method for predicting DNA N6-methyladenine sites in rice genome. *Front Genet* 2019;**10**:1071.
28. Wang HT, Xiao FH, Li GH, *et al.* Identification of DNA N^6-methyladenine sites by integration of sequence features. *Epigenetics Chromatin* 2020;**13**(1):8.
29. Huang Q, Zhang J, Wei L, *et al.* 6mA-RicePred: a method for identifying DNA N^6-methyladenine sites in the rice genome based on feature fusion. *Front Plant Sci* 2020;**11**:4.
30. Xu H, Hu R, Jia P, *et al.* 6mA-finder: a novel online tool for predicting DNA N6-methyladenine sites in genomes. *Bioinformatics* 2020;**36**(10):3257–9.
31. Cai J, Wang D, Chen R, *et al.* A bioinformatics tool for the prediction of DNA N6-Methyladenine modifications based on feature fusion and optimization protocol. *Front Bioeng Biotechnol* 2020;**8**:502.
32. Hasan MM, Basith S, Khatun MS, *et al.* Meta-i6mA: an interspecies predictor for identifying DNA N^6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. *Brief Bioinform* 2021;**22**(3):bbaa202.
33. Rehman MU, Chong KT. DNA6mA-MINT: DNA-6mA modification identification neural tool. *Gene* 2020;**11**(8):898.
34. Lv Z, Ding H, Wang L, *et al.* A convolutional neural network using dinucleotide one-hot encoder for identifying DNA N6-methyladenine sites in the rice genome. *Neurocomputing* 2021;**422**:214–21.
35. Xue T, Zhang S, Qiao H. i6mA-VC: a multi-classifier voting method for the computational identification of DNA N6-methyladenine sites. *Interdiscip Sci* 2021;**13**(3):413–25.
36. Huang Q, Zhou W, Guo F, *et al.* 6mA-Pred: identifying DNA N6-methyladenine sites based on deep learning. *PeerJ* 2021;**9**:e10813.
37. Li Z, Jiang H, Kong L, *et al.* Deep6mA: a deep learning framework for exploring similar patterns in DNA N6-methyladenine sites across different species. *PLoS Comput Biol* 2021;**17**(2):e1008767.
38. Rahman CR, Amin R, Shatabda S, *et al.* A convolution based computational approach towards DNA N6-methyladenine site identification and motif extraction in rice genome. *Sci Rep* 2021;**11**(1):10357.
39. Wang Y, Zhang P, Guo W, *et al.* A deep learning approach to automate whole-genome prediction of diverse epigenomic modifications in plants. *New Phytol* 2021;**232**(2):880–97.
40. Zeng R, Liao M. 6mAPred-MSFF: a deep learning model for predicting DNA N6-Methyladenine sites across species based on a multi-scale feature fusion mechanism. *Appl Sci* 2021;**11**(16):7731.
41. Hasan MM, Shoombuatong W, Kurata H, *et al.* Critical evaluation of web-based DNA N6-methyladenine site prediction tools. *Brief Funct Genomics* 2021;**20**(4):258–72.
42. Ryu JY, Lee MY, Lee JH, *et al.* DeepHIT: a deep learning framework for prediction of hERG-induced cardiotoxicity. *Bioinformatics* 2020;**36**(10):3049–55.
43. Wei L, Ye X, Xue Y, *et al.* ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. *Brief Bioinform* 2021;**22**(5):bbab041.
44. Medema MH, de Rond T, Moore BS. Mining genomes to illuminate the specialized chemistry of life. *Nat Rev Genet* 2021;**22**(9):553–71.
45. AlQuraishi M, Sorger PK. Differentiable biology: using deep learning for biophysics-based and data-driven modeling of molecular mechanisms. *Nat Methods* 2021;**18**(10):1169–80.
46. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;**28**(1):31–6.
47. Kipf TN, Welling M. *Semi-Supervised Classification with Graph Convolutional Networks*. *In: 5th International Conference on Learning Representations (ICLR)*. Toulon, France: OpenReview.net, 2017.
48. Shen LC, Liu Y, Song J, *et al.* SAResNet: self-attention residual network for predicting DNA-protein binding. *Brief Bioinform* 2021;**22**(5):bbab101.
49. Fu L, Niu B, Zhu Z, *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**(23):3150–2.
50. Zhang P, Wang Y, Chachar S, *et al.* eRice: a refined epigenomic platform for japonica and indica rice. *Plant Biotechnol J* 2020;**18**(8):1642.
51. Karimi M, Wu D, Wang Z, *et al.* DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* 2019;**35**(18):3329–38.
52. Zeng Y, Chen X, Luo Y, *et al.* Deep drug-target binding affinity prediction with multiple attention blocks. *Brief Bioinform* 2021;**22**(5):bbab117.
53. Yang Z, Zhong W, Zhao L, *et al.* ML-DTI: mutual learning mechanism for interpretable drug-target interaction prediction. *J Phys Chem Lett* 2021;**12**(17):4247–61.
54. O'Boyle NM, Banck M, James CA, *et al.* Open babel: an open chemical toolbox. *J Chem* 2011;**3**(1):1–14.
55. Kim S, Chen J, Cheng T, *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* 2019;**47**(D1):D1102–D9.
56. Abu-El-Haija S, Kapoor A, Perozzi B, *et al. N-gcn: Multi-scale graph convolution for semi-supervised node classification*. *In: Proceedings of The 35th Uncertainty in Artificial Intelligence Conference (UAI)*. *Tel Aviv*, Israel: PMLR, 2020, 841–51.

57. Chu Y, Wang X, Dai Q, *et al.* MDA-GCNFTG: identifying miRNA-disease associations based on graph convolutional networks via graph sampling through the feature and topology graph. *Brief Bioinform* 2021;**22**(6):bbab165.
58. David L, Thakkar A, Mercado R, *et al.* Molecular representations in AI-driven drug discovery: a review and practical guide. *J Chem* 2020;**12**(1):1–22.
59. Landrum G. RDKit documentation. *Release* 2013;**1**:4.
60. Li J, Zhang S, Liu T, *et al.* Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. *Bioinformatics* 2020;**36**(8):2538–46.
61. He K, Zhang X, Ren S, *et al. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, 2016, 770–8.
62. He K, Zhang X, Ren S, *et al. Identity mappings in deep residual networks. In: European conference on computer vision (ECCV)*. Amsterdam, Netherlands: Springer, 2016, 630–45.
63. Wang D, Zhang Z, Jiang Y, *et al.* DM3Loc: multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism. *Nucleic Acids Res* 2021;**49**(8):e46.
64. Bergstra J, Komer B, Eliasmith C, *et al.* Hyperopt: a python library for model selection and hyperparameter optimization. *Comput Sci Discov* 2015;**8**(1):014008.
65. You K, Long M, Wang J, *et al.* How does learning rate decay help modern neural networks? *CoRR* 2019; abs/1908.01878.
66. Li M, Zhang W. PHIAF: prediction of phage-host interactions with GAN-based data augmentation and sequence-based feature fusion. *Brief Bioinform* 2021;**00**(0):bbab348.
67. Zhang S, Zhao L, Zheng CH, *et al.* A feature-based approach to predict hot spots in protein-DNA binding interfaces. *Brief Bioinform* 2020;**21**(3):1038–46.
68. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;**10**(3):e0118432.
69. Chen Z, Zhao P, Li F, *et al.* iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief Bioinform* 2020;**21**(3):1047–57.
70. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**(86):2579–605.