DeepMethylation: A deep learning based framework with GloVe and transformer encoder for DNA methylation prediction

Zhe Wang Equal first author, [1] , **Sen Xiang** Corresp., Equal first author, [1] , **Chao Zhou** Corresp., [2] , **Qing Xu** [2]

[1] Wuhan University of Science and Technology, Unaffiliated, Wuhan, Hubei, China

[2] China Three Gorges University, Unaffiliated, Yichang, Hubei, China

Corresponding Authors: Sen Xiang, Chao Zhou
Email address: xiangsen@wust.edu.cn, zhouchao@ctgu.edu.cn

DNA methylation is a crucial topic in bioinformatics research. Traditional wet experiments are usually time-consuming and expensive. In contrast, machine learning offers an efficient and novel approach. In this study, we propose DeepMethylation, a novel methylation predictor with deep learning. Specifically, the DNA sequence is encoded with word embedding and GloVe in the first step. After that, dilated convolution and transformer encode modules are utilized to extract the features. Finally, full connection and softmax operations are applied to predict the methylation sites. The proposed model achieves an accuracy of 97.9% on the 5mC dataset, which outperforms state-of-the-art models. Furthermore, our predictor exhibits good generalization ability as it achieves an accuracy of 95.8% on the m1A dataset. To ease access for other researchers, our code is publicly available at https://github.com/sb111169/tf-5mc.

DeepMethylation: A deep learning based framework with GloVe and transformer encoder for DNA methylation prediction

Zhe Wang[1]**, Sen Xiang**[1]**, Chao Zhou**[2]**, and Qing Xu**[2]

[1]**School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan, Hubei, China**
[2]**School of Biology and Pharmacy, China Three Gorges University, Yichang, Hubei, China**

Corresponding author:
Sen Xiang and Chao Zhou

Email address: xiangsen@wust.edu.cn, zhouchao@ctgu.edu.cn

ABSTRACT

DNA methylation is a crucial topic in bioinformatics research. Traditional wet experiments are usually time-consuming and expensive. In contrast, machine learning offers an efficient and novel approach. In this study, we propose DeepMethylation, a novel methylation predictor with deep learning. Specifically, the DNA sequence is encoded with word embedding and GloVe in the first step. After that, dilated convolution and transformer encode modules are utilized to extract the features. Finally, full connection and softmax operations are applied to predict the methylation sites. The proposed model achieves an accuracy of 97.9% on the 5mC dataset, which outperforms state-of-the-art models. Furthermore, our predictor exhibits good generalization ability as it achieves an accuracy of 95.8% on the m1A dataset. To ease access for other researchers, our code is publicly available at https://github.com/sb111169/tf-5mc.

Subjects Bioinformatics, Computational Science, Data Mining and Machine Learning
Keywords DNA methylation, word vector model, deep learning, transformer, site prediction

INTRODUCTION

Epigenetics is first introduced to study the heritable changes in the regulation of gene expression without altering the nucleotide sequence of DNA. Advancements in life sciences have led to constant updates to the definition of epigenetics. Researchers have discovered various epigenetic mechanisms, including protein acetylation and methylation(Zhang et al., 2020a). Currently, N6-methyladenine (6mA), N4-methylcytosine (4mC) and 5-methylcytosine (5mC) are the three most widely studied types of DNA methylation. Take 5mC as an example, it commonly appears on the fifth carbon atom of cytosine in the DNA sequence's CpG dinucleotides. DNA methyltransferase transfers the methyl (-CH3) group from S-AdenosylMethionine (SAM) to the fifth carbon atom of cytosine(Adampourezare et al., 2021).

Studies have indicated the possible negative impact on organisms of abnormal DNA methylation. Firstly, DNA methylation can affect the level of gene expression, and even lead to gene silencing or abnormal expression(Ehrlich, 2003). For example, DNA methylation can change the conformation of chromatin, thus affecting chromatin accessibility and gene expression. In addition, the risk of gene mutations is positively correlated with DNA methylation(De Bont and Van Larebeke, 2004). Methylation sites are prone to be damaged in the process of replication and repair of DNA. If they are not repaired correctly, it may lead to loss of DNA or accumulation of mutations. Moreover, the same is true of the occurrence and development of cancer(Xu et al., 2011; Chowdhury et al., 2011; Lu et al., 2012; Koivunen et al., 2012). Some cancer cells have aberrant methylation of genes involved in important cellular life processes such as cell growth, differentiation and apoptosis, suggesting that DNA methylation may promote tumor initiation and progression. For instance, mutations in IDH1/2 produce the oncogenic metabolite 2-HG, which results in increased DNA methylation at the cellular level. This alteration affects gene expression and leads to cancer. Finally, embryonic development and adult diseases are also associated with DNA methylation(Jin et al., 2008; Tatton-Brown et al., 2014; Baets et al., 2015). DNA methylation plays an important role in embryonic development, and abnormal methylation may cause birth defects or abnormal development. The status of three functional protein families in the epigenetic system (write, reader, eraser), and their associated genes' genetic variation can cause diseases (e.g., autism, blood disease) by affecting overall cell-level epigenetics. Therefore, DNA methylation plays an important role in gene expression regulation and chromatin structure variation, and the detection of methylation is of great importance.

Current methods for methylation detection include wet experiments, traditional machine learning methods, and deep learning methods. Wet experiments conduct molecular biology tests to distinguish between methylation and demethylation in DNA samples. This typically involves bisulfite treatment(Smallwood et al., 2014; Kernaleguen et al., 2018), enzymatic digestion, and chromatin immunoprecipitation. Following bisulfite treatment, methylated cytosine is oxidized and transformed to unmethylated uracil, whereas unmethylated cytosine remains unchanged, and the difference indicates methylation.

Traditional machine learning methods generally consist of three key steps: data processing, feature extraction, and classification, which are all designed based on the experience of the researchers. Commonly-used features include physical, statistical, and sequence annotation features such as base frequencies, G+C content, length, repetitive sequences, RNA elements, and protein binding sites (Fang et al., 2006; Zhang et al., 2015). Based on the features, classification algorithms like logistic regression, support vector machines, or decision trees are used to identify the methylation sites.

In contrast, deep learning methods are more straightforward. Instead of manually specifying the feature extractor and classifier, researchers only need to design the deep neural networks, which automatically extract features and predict methylation results from DNA sequences. Furthermore, driven by datasets with a large number of samples, deep learning can extract more essential features than manually designed ones. For instance, the DNA module and the CpG module of the DeepCpG model(Angermueller et al., 2017) can predict the relationship between DNA sequences and their methylation status, as well as the relationship between adjacent CpG sites within a single cell or across cells.

In general, wet experiments achieve high accuracy, but they only predict a small number of DNA methylation sites. In addition, conducting wet experiments requires not only great cost and time, but also professional knowledge in biology, and these factors make it difficult to be widely applied. Traditional machine learning needs specified feature design, which also requires professional experience and extensive tests to find good feature descriptors. In contrast, deep learning methods can automatically learn the most relevant features without specifying them in advance, and can handle large datasets and high-dimensional data.

However, although deep learning methods provide new insights to detect DNA methylation, they still face challenges. On the one hand, convolutional neural networks (CNNs) can extract features for DNA, but they are not sensitive to 1D sequential data. On the other hand, recurrent neural networks (RNNs) are more suitable for feature extraction of sequential signals, but they do not perform well in learning remote relationship. Moreover, conducting large-scale parallel computation is challenging due to RNN's structure. In addition, the current DNA encoding methods, one-hot and word embedding, emphasize local information and ignore global relationship.

To solve these problems, in this paper, we propose DeepMethylation, a novel deep-learning based scheme to predict DNA methylation sites. The contribution of this paper is as the following. Firstly, with word embedding and GloVe, we propose a novel DNA encoding method. This new representation format improves the ability in modeling the relationship between DNA sub-sequences. Secondly, dilated convolution and transform encoder are incorporated to better extract both local and global features, especially the relationship between DNA sequences far from each other. Last but not least, dense full connections are used to predict the methylation statue of each site. Experimental results demonstrate that the accuracy of the proposed method reaches 97.9%, which outperforms other state-of-the-art methods.

RELATED WORKS

Wet experiments

Genome-wide single nucleotide resolution (GWGSR) typically requires wet experiments to be realized. Currently, the main approaches for achieving GWGSR include whole-genome bisulfite sequencing (WGBS), reduced representation bisulfite sequencing (RRBS), and DNA methylation chip.

WGBS (Smallwood et al., 2014; Kernaleguen et al., 2018) is a high-resolution and comprehensive method for full-genome sequencing via bisulfite treatment, which converts unmethylated cytosine to uracil, but does not convert methylated cytosine. Methylation status of individual cytosines can be determined at the single nucleotide level by comparing the DNA sequences with and without bisulfite treatment. RRBS(Guo et al., 2013; Farlik et al., 2015; Hou et al., 2016) is a cost-effective alternative to WGBS and involves sequencing the CpG-rich subset of the genome. RRBS reduces the requirement for sequence depth to cover the entire genome and still provides single nucleotide resolution at CpG sites. DNA methylation chips(Morris et al., 2014) represent microarray-based platforms that simultaneously detect DNA methylation levels among thousands of CpG sites in the genome. These chips contain probes that are specific to methylated or unmethylated CpG sites, and the intensity of the signal from each probe indicates the site's methylation level.

Although wet experiments produce accurate prediction results, it needs great financial cost and time, as well as professional biology knowledge, which is inefficient in implementation.

Traditional machine learning methods

With the rapid advancement in automatic DNA sequencing technology, huge amount of DNA sequences are obtained, promoting the analysis of DNA data. Traditional machine learning methods involve two steps. Firstly, manually designed DNA features are proposed. After that, with these features, machine learning classification algorithms are utilized to predict the methylation. Stevens et al.(Stevens et al., 2013) integrated the features from chromatin immunoprecipitation sequencing and methylation-sensitive restriction enzyme sequencing, and predicted the methylation status of CpG sites in the human genome by using a conditional random field model. Zhang et al. (Zhang et al., 2015)utilized various features, including methylation markers, genomic locations, and regulatory factors, to design a methylation prediction model with a random forest classifier. Fang et al.(Fang et al., 2006) developed a CpG island methylation prediction tool called MethCGI using CpG island data from the human brain. This model takes input features such as CpG ratio, GC content, TpG frequency and transcription factor binding site distribution, and employs a support vector machine as a classifier.

Machine learning methods have demonstrated higher efficiency and lower costs than wet experiments. However, the performance of machine learning models is limited by the manual selection of feature descriptors and classifiers, which relies on the experience of the researchers.

Deep learning methods

In recent years, with the rapid development of neural networks, deep learning methods have been applied to DNA methylation prediction(Routhier and Mozziconacci, 2022). Deep learning automatically extracts features and is free from tedious feature engineering, allowing an end-to-end model to be constructed for feature extraction and classification. Deep learning methods, including convolutional neural networks (CNN) and recurrent neural networks (RNN), have been proven to perform well in predicting DNA methylation sites.

Angermueller et al. proposed the DeepCpG model(Angermueller et al., 2017). The model consists of DNA module, CpG module and joint module. The DNA module involves two convolutional layers and a pooling layer to identify correlations between DNA sequence patterns and methylation status. The CpG module employs a bidirectional gated recurrent network to identify correlations between adjacent CpG sites. The joint module learns the interaction between the DNA and CpG modules to predict the methylation status in all cells. Tian et al. proposed MRCNN(Tian et al., 2019), which used the correlation between DNA sequence patterns and methylation levels to predict the methylation of the CpG site at single base resolution. The model used one-hot encoding, convolution, pooling and fully connected layers to output the predicted value. With a continuous loss function, MRCNN achieves smooth regression of methylation values, and produces more accurate results than the DeepCpG model. Zhou et al. built a RNN-based DNA methylation prediction model(Zhou, 2020), this model first converts the raw DNA sequence into matrix data through one-hot encoding, and then sends it to the RNN model for feature extraction and methylation prediction. The results have shown that RNNs are more suitable for handling sequence data and extracting hidden temporal features from the sequence than CNNs. Cheng et al. proposed iPromoter-5mC(Cheng et al., 2021), believing that DNA chemical properties can affect its genetic traits. To address this issue, they combined one-hot encoding with deoxyribonucleic acid nucleotide properties and their frequency (DPF) to generate a composite feature set. They then used a deep neural network to process the composite feature set for identifying methylation modification sites in promoters. Tran et al. considered that the DNA sequence can be regarded as a distinct linguistic system, and they proposed an efficient encoding method to identify 5-methylcytosine sites. By embedding k-mers, they transformed the DNA sequence into ‘sentences’, and then generate the feature vector of the DNA sequence(Tran et al., 2021) with k-mers representation. Then, the feature vectors were separately sent to xgboost, random forest, deep forest and deep feedforward neural network. The final results showed that the performance of this model was better than iPromoter-5mC.

In general, deep neural networks have better learning abilities than traditional learning methods, and thus produce more accurate results. Nevertheless, CNNs and RNNs still encounter challenges in encoding feature representations and efficiently extracting global long-distance features, and further research is desired.

MATERIALS AND METHODS

The Overall Framework

As shown in figure 1, the proposed DeepMethylation has 3 modules, which are data processing module, feature extraction module and classification module. First, in data processing module, the one-dimensional DNA sequence is segmented and converted to a 39 × 300 matrix with word embedding and GloVe. After that, the feature extraction module utilizes transformer encoder and dilated convolution to extract global and local features. Finally, with the extracted features, the classification module predicts the methylation state of each site of the DNA sequence.

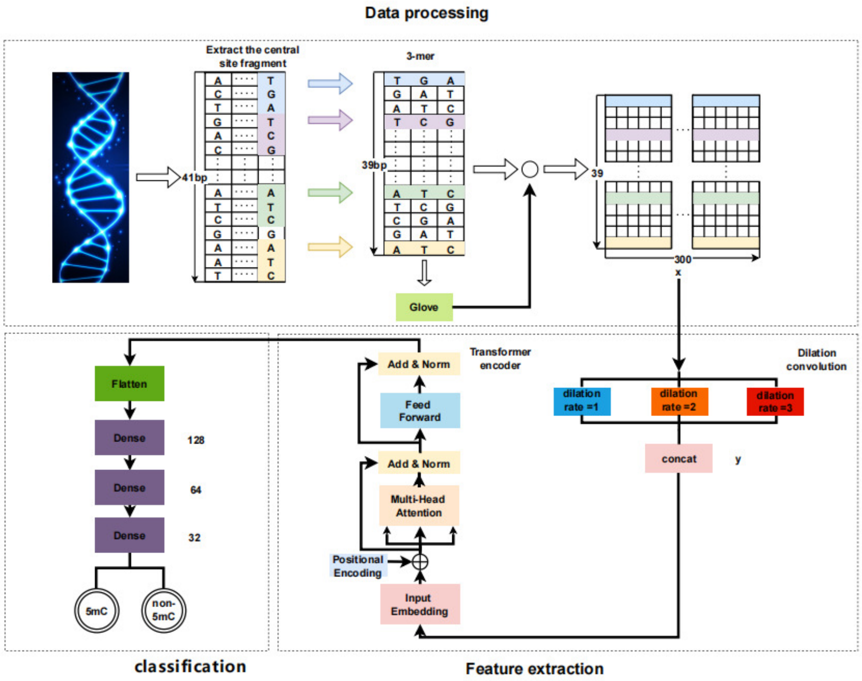


Figure 1. The overall framework of DeepMethylation.

Data processing

As a long sequence, DNA is not conducive to presenting the relationship among different fragments. Therefore, the first step is to convert one-dimensional DNA sequences to a group of short fragments. Although one-hot encoding(Abbas et al., 2021) can represent each base of DNA as a binary bit, it cannot provide the sequence orders or measure the distance(Huang et al., 2021) between related words. In this paper, word embedding and GloVe algorithm are used to better model the relationship in DNA sequences.

By following the rule of WGBS, the golden standard of methylation detection, DNA sequences are cropped into 41 bp segments. As shown in figure 2, a 3-bp window slides over a segment and produces a series of 3-mer sub-sequences. As a result, a 41bp DNA sequence is converted to a 39×3 3-mer matrix.

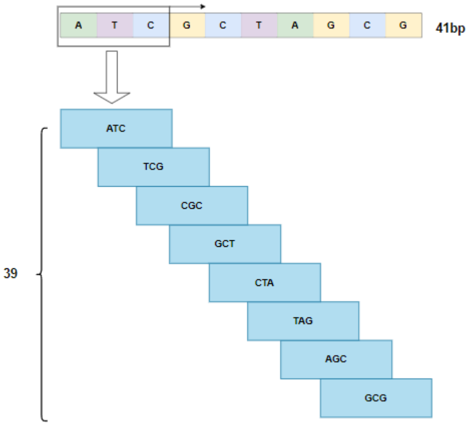
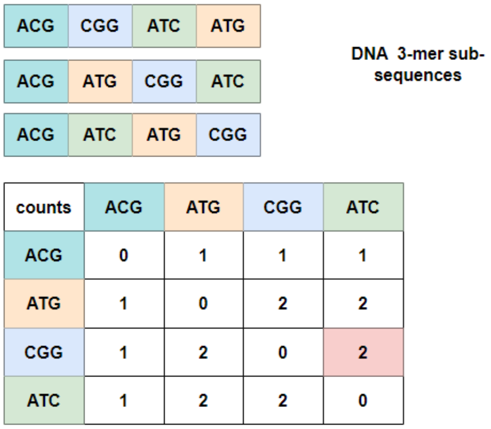


Figure 2. 3-mer sub-sequences.



counts	ACG	ATG	CGG	ATC
ACG	0	1	1	1
ATG	1	0	2	2
CGG	1	2	0	2
ATC	1	2	2	0

Figure 3. Co-occurrence matrix.

To explore the relationship between these 3-mer fragments, GloVe(JeffreyPennington and Manning, 2014; Wang et al., 2022), a word embedding model based on global vectors, is utilized. It first checks the context of neighboring 3-mer fragments and obtains a co-occurrence matrix. Figure 3 depicts an example of the co-occurrence matrix for three four-word sentences. Take the combination of 'CGG-ATC' as an example, it happens twice that 'CGG' appears before 'ATC', and the corresponding intersection with the row index 'CGG' and the column index 'ATC' is valued at two, which is marked in pink in the co-occurrence matrix for better illustration. Mathematically, the co-occurrence matrix is notated as X, and $X_{i,j}$ represents the frequency of word j appearing after i. I this way, the example in figure 3 can also be noted as $X_{CGG,ATC} = 2$. Moreover, it is noteworthy that the matrix is symmetric about the diagonal line, and the elements in the upper-right of the matrix are computed and copied to the lower-left.

In figure 3, the co-occurrence matrix only indicates the relationship of three sub-sequences, and in order to model the relationship for all sub-sequences, GloVe algorithm traverses the entire corpus and derives a global word vector dictionary through inner product operation and translation transformation of words (Cochez et al., 2017; Liu et al., 2019a), which makes the mapping values equal or approximate to the co-occurrence probability of words. To be specific, an energy function J is defined as

$$J = \sum_{i,j=1}^{N} f(X_{i,j}) \left[V_i^T \widetilde{V}_j + b_i + b_j - log(X_{i,j}) \right]^2 \tag{1}$$

where b_i and b_j are offsets, and N is the total number of words. V_i represent the word vector in the global dictionary to be obtained, $\widetilde{V}_j$ is the separate context vector that help solving V_i. Since J is a convex function, V_i can be solve via optimization algorithms such as gradient descent. In addition, the weighting factor $f(X_{i,j})$ is defined as

$$f(X_{i,j}) = \begin{cases} \left[\frac{X_{i,j}}{T_X} \right]^{\alpha} & \text{if } X_{i,j} < T_X \\ 1 & otherwise \end{cases} \tag{2}$$

where T_X is a threshold. With truncation and non-linear mapping, the weighting factor can retain crucial information in the co-occurrence count while also eliminating noise and irregular co-occurrence. In very special cases, $f(X_{i,j})$ equals to 1 only when two words are semantically similar and locate closely to each together in the vector space. α is set to 0.75, which enables the model to achieves quite good performances as has been proved in (JeffreyPennington and Manning, 2014).

In implementation, the length of V_i is set to 300 for each valid vector word. Once all word vectors V_i are obtained, each 3-mer word can be represented by the corresponding word vector. As a result, the 39 3-mer words in figure 1 can be presented with a 39×300 encoding vector matrix.

Feature extraction

After data processing, the 39×300 word vector matrix is used for feature extraction. To be specific, this matrix is regarded as a word vector embedding layer that utilized as input for the feature extraction module, which utilizes dilated convolution and transformer encoder as shown in figure 1.

On the one hand, to enlarge the receptive field while keep low computational complexity(Liu et al., 2019b; Yuan et al., 2019), dilated convolution is utilized. As shown in figure 4, in dilated convolution, the filter is expanded by inserting zeros between its values. This effectively increases the receptive field of the filter without increasing the number of parameters, allowing it to capture larger spatial structures and longer-term dependencies of the input. In this study, three branches with dilation rates of 1, 2, and 3 are used, followed by features concatenation, producing feature of contextual information at different scales.

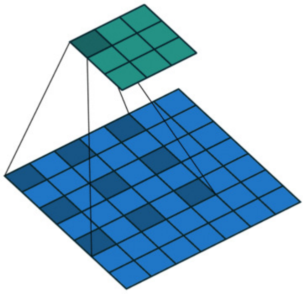


Figure 4. Dilated convolution.

On the other hand, followed by dilated convolution, a transformer encoder is used to extract the global relationship in the spliced features and the long-term dependency relationship between elements in the sequence(Khan et al., 2022). As shown in figure 5, based on the transformer encoder, which consists of input embedding, multi-head attention, add&norm, and feed forward, we incorporate positional encoding into the module. Positional encoding is an important property for sequential signals, take the 39 DNA fragments shown in figure 2 as an example, if their positions or arrangement orders are changed, they will form a new DNA sequence that is totally different. Therefore, with positional encoding, word orders can be introduced to distinguish between DNA sequences.

Another important mechanism in the transformer encoder is the multi-head attention (MHA), which computes the relative importance between different positions in the input sequence so as to provide better input feature representation for the subsequent feed forward network. Figure 6 depicits the framework of MHA. The input features, which are in the form of 3D tensors, are copied multiple times. For each feature, a weighting factor is calculated with the self-attention mechanism, with which the weighted summation of the input features are calculated. MHA can map the input features to multiple sub-spaces, and improves the model's understanding of the input sequence with feature extraction, attention calculation and feature concatenation. Furthermore, each head in MHA works independently, thus expanding the decision space of the model and enabling better decisions while mitigating over-fitting.

Finally, after the operation of 'addition and normalization (Add&Norm)', the transformer produces the features of the gene, which are used for classification.

Classification

As shown in figure 2, the features extracted by the encoder are finally sent to the classifier to predict the sites of mathylation. The classifier has three fully connected layers with dropout. The dimensions of the three fully connected layer are 128, 64 and 32, respectively. Finally, a Sigmoid activation function is used to report whether a site is 5mc or non-5mc. In addition, the categorical cross-entropy loss is adopted to train the network.

RESULTS AND ANALYSIS

Dataset

In learning-based methods, the dataset is of fundamental importance. In this study, we use the Cancer Cell Line Encyclopedia (CCLE) dataset proposed by Zhang et al.(Zhang et al., 2020b), where 5mC

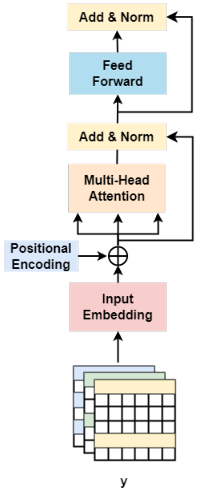


Figure 5. Transformer Encoder.

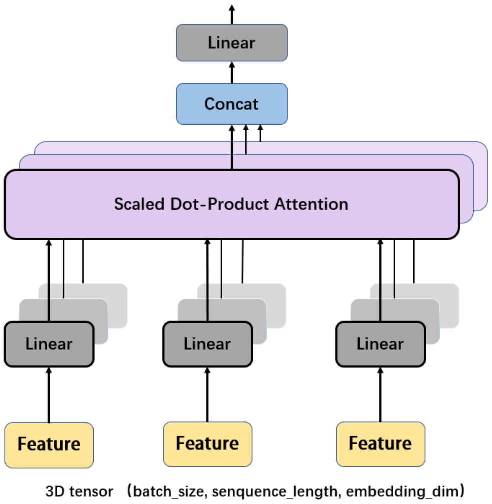

Linear
Concat
Scaled Dot-Product Attention
Linear
Linear
Linear
Feature
Feature
Feature
3D tensor（batch_size, senquence_length, embedding_dim)

Figure 6. Multi-head Attention

modification sites of various cancer cell lines are processed by simplified RRBS experiment. Especailly, we focus on investigating the distribution of 5mC sites in small cell lung cancer (SCLC)(Barretina et al., 2012; Li et al., 2019). DNA fragments with ‘C’ locating in the center are extracted and notated as

$$E_{(\delta)}\,(C) = \,E_{-(\delta)}\,E_{-(\delta-1)}\,\cdots\,E_{-(1)}\,C\,E_{+(1)}\,\cdots\,E_{+(\delta-1)}\,E_{+(\delta)} \tag{3}$$

where for each site $E \in \{A,T,G,C\}$. In implementation, by following the rule of WGBS, δ is set to 20, and each fragment has 41 sites. In this way, a total of 93000 DNA fragments are obtained, including 65000 methylation-positive samples and 865000 negative ones. As shown in table 1 and figure 7, the ratio between the negative and positive samples is about 13.3, which coincides with the distribution of 5mC in real cases.

The experiment is conducted on a server with an Intel(R) Core(TM) i9-10900F CPU, a 64GB RAM, and an NVIDIA GeForce RTX 3090 GPU. The software is programmed with Python 3.7, Keras-nightly 2.8, and tf-nightly-gpu 2.8.0.

Table 1. The information of the experimental datasets.

Dataset	Positive Sample	Negative Sample
Training Dataset	52000	692000
Testing Dataset	13000	173000
Total	65000	865000

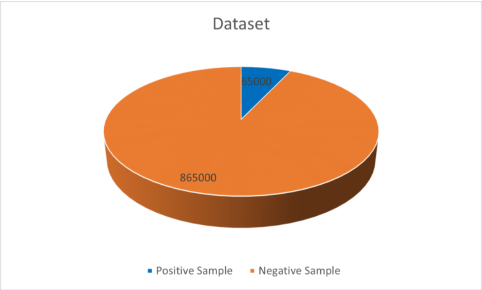


Figure 7. Proportion of positive and negative samples

Performance Evaluation

The model is trained and tested with the aforementioned dataset. According to the test results, e.g. the numbers of true negative (TN), false negative (FN), true positive (TP), and false positive (FP) samples, the following indexes are computed to evaluate the performance of the model.

- Sensitivity (Sen) refers to the ratio of correctly predicted positive samples to all positive samples.

$$Sen = \frac{TP}{TP+FN} \tag{4}$$

- Specificity (Spe) refers to the ratio of correctly predicted negative samples to all negative samples.

$$Spe = \frac{TN}{TN+FP} \tag{5}$$

- Accuracy (Acc) refers to the ratio of correctly classified samples, both positive and negative, to all tested samples.

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

- Matthews Correlation Coefficient (Mcc) considers the joint relationship between TP, TN, FP and FN, and comprehensively evaluates the consistency between the predicted results and the ground truth.

$$Mcc = \frac{TP\times(TN)-FP\times(FN)}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \tag{7}$$

- Area Under the Curve (AUC) compares the performance of different models by calculating the area under the Receiver Operating Characteristic (ROC) curve, and larger value indicates higher degree of authenticity.

Performance Comparison with SOTA Methods

Three state-of-the-art (SOTA) methylation prediction methods, iPromoter-5mC(Cheng et al., 2021), 5mC-Pred(Tran et al., 2021) and BiLSTM-5mC(Zhang et al., 2020b), are compared with our model. In order to make a fair comparison, all models are trained with the aforementioned dataset, and are subjected to 5-fold cross-validation. Table 2 presents the technique features, including encoding, feature extraction, and classification of the methods.

Table 2. Summary of existing tools for 5mC sites prediction in genome-wide DNA promoters.

Method	Encoding	Feature Extraction and Classification
iPromoter-5mC	One-hot	Deep neural network
5mC-Pred	K-mers	XGBoost
BiLSTM-5mC	One-hot and NPF	BiLSTM
Our model	GloVe	Digital convolution and transformer encoder

As shown in figures 8-12, our model performs the best in terms of Spe, Acc, Mcc, and Auc, indicating that our model can get more essential features and the classifier is also more accurate. Our model adopts encoding technique, including word embedding and GloVe, and transformer feature extraction, as well as dilated convolution. These techniques improves the ability in modeling the relationship among sub-sequences, which also benefits the accurate classification of methylation for the gene sites.

We also noticed that, in terms of ‘Sen’, the proposed framework is slightly lower than iPromoter-5mC and 5mC-Pred, the reason is that the our method focus on making reliable predictions, or in other words, our model is trend to classify a positive sample as ‘negative’ if it is not that confident. As a result, ‘TP’ becomes slightly smaller and ‘FN’ is larger than the ground truth. Although this reduces the value of ‘Sen’, it provides more reliable judgement for ‘TP’. On the other hand, it should be noticed that, in terms of the overall accuracy ‘Acc’ that takes all tested samples involved, the proposed method reaches 0.979, which is about 5% higher than the sub-optimal method BiLSTM-5mC.

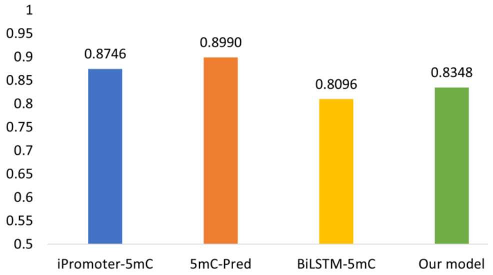


Figure 8. Sen comparison.

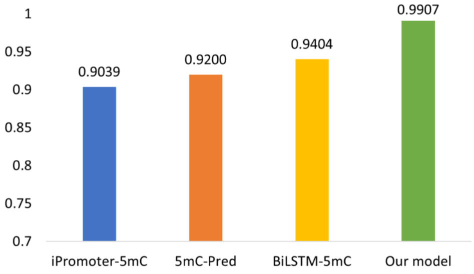


Figure 9. Spe comparison.

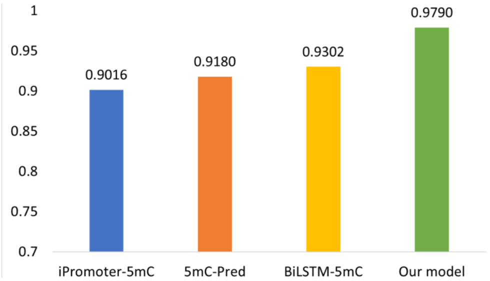


Figure 10. Acc comparison.

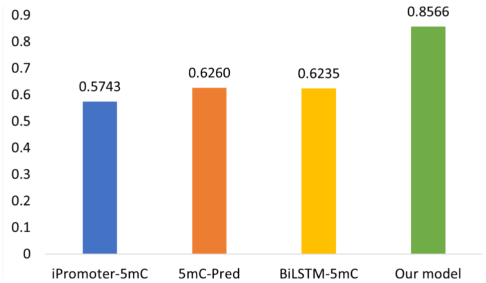


Figure 11. Mcc comparison.

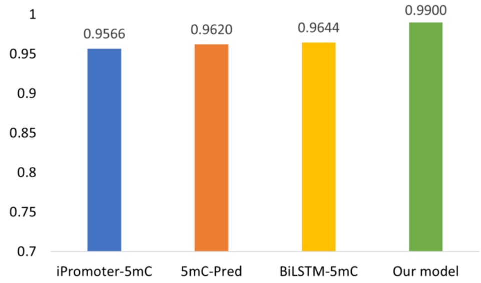


Figure 12. Auc comparison.

Influence of encoding methods

Encoding methods have a significant impact on the model's performance. In addition to the word embedding and GloVe encoding used in this paper, one-hot encoding(Vinyals et al., 2016) is also widely utilized. To verify the superiority of GloVe, we replace the encoding method in figure 1 with one-hot encoding, and compare the performance with the five quality indexes as shown in figure 13. It can be noticed that, both methods produce satisfactory results, but GloVe still performs better than one-hot encoding.

To be specific, for Spe, Acc and Auc, the performance of the methods are similar with index values above 0.97. For the other two indexes, Sen and Mcc, GloVe encoding achieves significant performance improvement over one-hot encoding. The reason is that one-hot encoding only provides the simplest mapping of the four bases A, T, C, and G, resulting in low-dimensional representation of DNA, while GloVe incroprates sliced DNA fragments, and thus better represents the relationship among sub-sequences. Therefore, GloVe encoding exhibits better ability to identify positive examples, as well as higher correlation between the predictions and the ground truth.

Influence of feature extraction methods

In addition to encoding, feature extraction methods also greatly affect the methylation detection results. The long short term memory (LSTM)(Yu et al., 2019) is widely used to extract features for 1D sequence, so a comparison is made between LSTM and the proposed Transform encoder. As shown in figure 14, it can be seen that the transformer encoder achieves better performance than LSTM in terms of Sen, Mcc and Auc. The reason is that, as a recurrent neural network, LSTM relies on memory units to transmit information when dealing with long sequences. However, as the sequence grows longer, the information

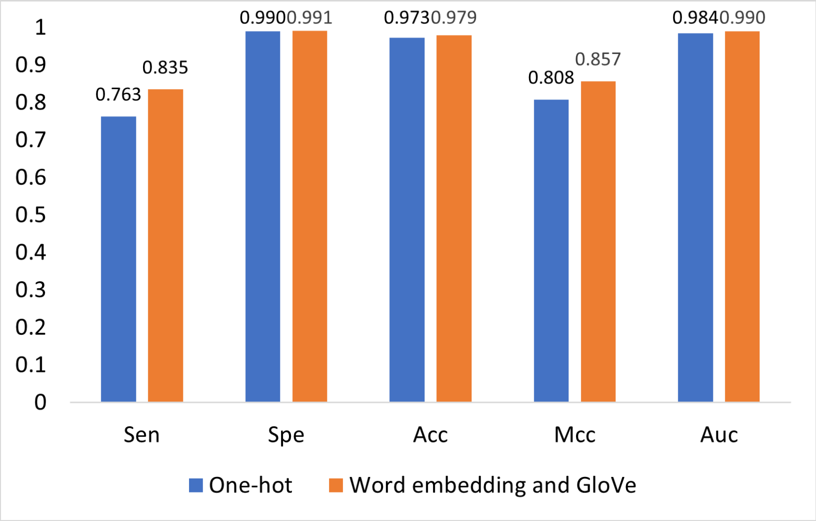


Figure 13. Performance comparison of feature encoding methods for the prediction of 5mC sites.

transmission becomes weaker in the network, which impairs the ability of long-term modeling. In comparison, the transformer encoder utilizes the multi-head attention mechanism, which directly model the relationship between input signals without relying on the context, and thus better deal with long sequence data.

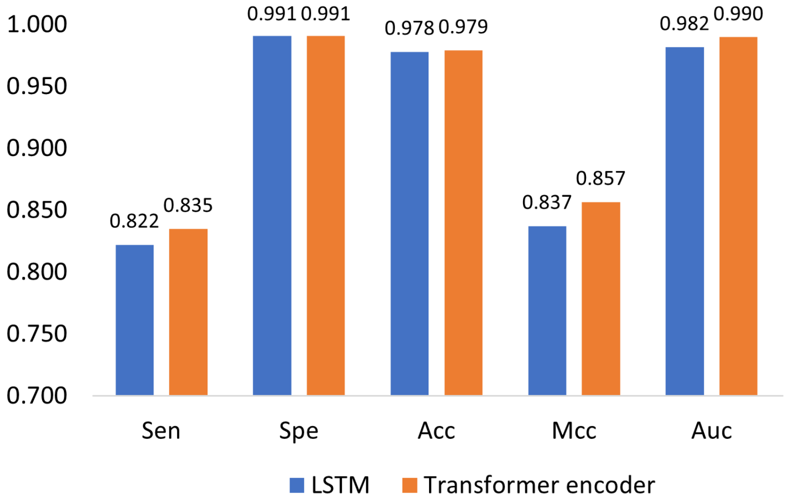


Figure 14. Performance Evaluation of Different Feature Extraction Methods.

Generality on m1A data

We also test the generalization performance of the proposed method when extending to other types of DNA data. For example, in figures 15-16, the proposed model is tested with m1A data of the EMDLP dataset (Wang et al., 2022). It can be noticed that, the network converges after 12 epoches, and the accuracy reaches 95.8%. This demonstrates that the proposed method has good generation ability and is promising to be applied to different DNA data.

CONCLUSIONS

After analyzing existing research and extensively comparing experimental performance, the proposed DeepMethylation is proved to be an effective method for identifying DNA methylation sites. Word embedding and GloVe can effectively describe the features of DNA sequences and reveal hidden relationships

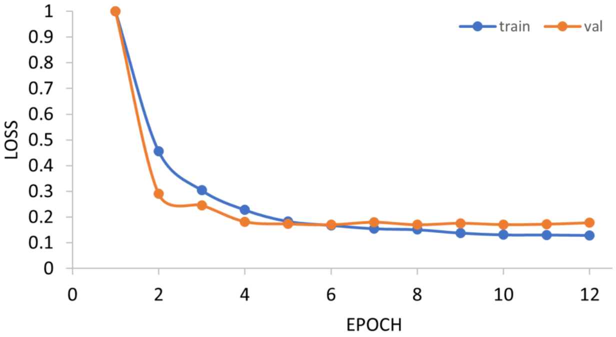


Figure 15. Training Progress of training and testing data sets.

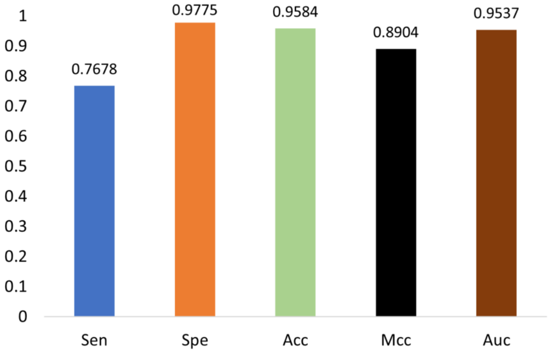


Figure 16. Performance on m1A dataset.

between sub-sequences. In addition, the feature extraction module, including the transformer encoder and the dilated convolution, can better extract local and global features of DNA sequences. Experiments demonstrated that the proposed framework, as well as the specified tools, achieves accurate methylation detection, with the accuracy reaching 97.9%, and can be applied to m1A data. In the future, we will consider using this model for the recognition of 4mC and 6mA.

Although excellent research has been done by previous researchers in this field and they have provided theoretical and experimental support for our research, there is still room for improvement, such as developing unsupervised learning methylation detection approaches without the ground truth labels, which are commonly obtained via expensive and labor-intensive wet experiments.

ACKNOWLEDGMENTS

Funding This work was supported by the Natural Science Foundation of Hubei Province (Grant 2022CFB349). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures The following grant information was disclosed by the authors: Natural Science Foundation of Hubei Province grant: 2022CFB349.

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

Author Contributions

Zhe Wang conceived and designed the experiments, performed the experiments, prepared figures and tables, authored or reviewed drafts of the paper.

Sen Xiang conceived and designed the experiments, prepared figures and tables, authored or reviewed drafts of the paper, and approved the final draft.

Chao Zhou conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Qing Xu analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data availability

The following information was supplied regarding data availability:

Raw data is available at GitHub: https://github.com/sb111169/tf-5mc/tree/main/shuju

Code is also available at GitHub: https://github.com/sb111169/tf-5mc/tree/main

REFERENCES

Abbas, Z., Tayara, H., and Chong, K. T. (2021). 4mcpred-cnn—prediction of dna n4-methylcytosine in the mouse genome using a convolutional neural network. *Genes*, 12(2):296.

Adampourezare, M., Dehghan, G., Hasanzadeh, M., and Feizi, M.-A. H. (2021). Application of lateral flow and microfluidic bio-assay and biosensing towards identification of dna-methylation and cancer detection: Recent progress and challenges in biomedicine. *Biomedicine & Pharmacotherapy*, 141:111845.

Angermueller, C., Lee, H. J., Reik, W., and Stegle, O. (2017). Deepcpg: accurate prediction of single-cell dna methylation states using deep learning. *Genome biology*, 18(1):1–13.

Baets, J., Duan, X., Wu, Y., Smith, G., Seeley, W. W., Mademan, I., McGrath, N. M., Beadell, N. C., Khoury, J., Botuyan, M.-V., et al. (2015). Defects of mutant dnmt1 are linked to a spectrum of neurological disorders. *Brain*, 138(4):845–861.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607.

Cheng, X., Wang, J., Li, Q., and Liu, T. (2021). Bilstm-5mc: a bidirectional long short-term memory-based approach for predicting 5-methylcytosine sites in genome-wide dna promoters. *Molecules*, 26(24):7414.

Chowdhury, R., Yeoh, K. K., Tian, Y.-M., Hillringhaus, L., Bagg, E. A., Rose, N. R., Leung, I. K., Li, X. S., Woon, E. C., Yang, M., et al. (2011). The oncometabolite 2-hydroxyglutarate inhibits histone lysine demethylases. *EMBO reports*, 12(5):463–469.

Cochez, M., Ristoski, P., Ponzetto, S. P., and Paulheim, H. (2017). Global rdf vector space embeddings. In *The Semantic Web–ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part I 16*, pages 190–207. Springer.

De Bont, R. and Van Larebeke, N. (2004). Endogenous dna damage in humans: a review of quantitative data. *Mutagenesis*, 19(3):169–185.

Ehrlich, M. (2003). Expression of various genes is controlled by dna methylation during mammalian development. *Journal of cellular biochemistry*, 88(5):899–910.

Fang, F., Fan, S., Zhang, X., and Zhang, M. Q. (2006). Predicting methylation status of cpg islands in the human brain. *Bioinformatics*, 22(18):2204–2209.

Farlik, M., Sheffield, N. C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., and Bock, C. (2015). Single-cell dna methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell reports*, 10(8):1386–1397.

Guo, H., Zhu, P., Wu, X., Li, X., Wen, L., and Tang, F. (2013). Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome research*, 23(12):2126–2135.

Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., Wu, X., Wen, L., Tang, F., Huang, Y., et al. (2016). Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell research*, 26(3):304–319.

Huang, Q., Zhou, W., Guo, F., Xu, L., and Zhang, L. (2021). 6ma-pred: identifying dna n6-methyladenine sites based on deep learning. *PeerJ*, 9:e10813.

JeffreyPennington, R. and Manning, C. (2014). Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing. Citeseer*.

Jin, B., Tao, Q., Peng, J., Soo, H. M., Wu, W., Ying, J., Fields, C. R., Delmas, A. L., Liu, X., Qiu, J., et al. (2008). Dna methyltransferase 3b (dnmt3b) mutations in icf syndrome lead to altered epigenetic modifications and aberrant expression of genes regulating development, neurogenesis and immune function. *Human molecular genetics*, 17(5):690–709.

Kernaleguen, M., Daviaud, C., Shen, Y., Bonnet, E., Renault, V., Deleuze, J.-F., Mauger, F., and Tost, J. (2018). Whole-genome bisulfite sequencing for the analysis of genome-wide dna methylation and hydroxymethylation patterns at single-nucleotide resolution. *Epigenome Editing: Methods and Protocols*, pages 311–349.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. (2022). Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41.

Koivunen, P., Lee, S., Duncan, C. G., Lopez, G., Lu, G., Ramkissoon, S., Losman, J. A., Joensuu, P., Bergmann, U., Gross, S., et al. (2012). Transformation by the (r)-enantiomer of 2-hydroxyglutarate linked to egln activation. *Nature*, 483(7390):484–488.

Li, H., Ning, S., Ghandi, M., Kryukov, G. V., Gopal, S., Deik, A., Souza, A., Pierce, K., Keskula, P., Hernandez, D., et al. (2019). The landscape of cancer cell line metabolism. *Nature medicine*, 25(5):850–860.

Liu, X.-Q., Li, B.-X., Zeng, G.-R., Liu, Q.-Y., and Ai, D.-M. (2019a). Prediction of long non-coding rnas based on deep learning. *Genes*, 10(4):273.

Liu, Y., Dong, H., Wang, X., and Han, S. (2019b). Time series prediction based on temporal convolutional network. In *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, pages 300–305. IEEE.

Lu, C., Ward, P. S., Kapoor, G. S., Rohle, D., Turcan, S., Abdel-Wahab, O., Edwards, C. R., Khanin, R., Figueroa, M. E., Melnick, A., et al. (2012). Idh mutation impairs histone demethylation and results in a block to cell differentiation. *Nature*, 483(7390):474–478.

Morris, T. J., Butcher, L. M., Feber, A., Teschendorff, A. E., Chakravarthy, A. R., Wojdacz, T. K., and Beck, S. (2014). Champ: 450k chip analysis methylation pipeline. *Bioinformatics*, 30(3):428–430.

Routhier, E. and Mozziconacci, J. (2022). Genomics enters the deep learning era. *PeerJ*, 10:e13613.

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods*, 11(8):817–820.

Stevens, M., Cheng, J. B., Li, D., Xie, M., Hong, C., Maire, C. L., Ligon, K. L., Hirst, M., Marra, M. A., Costello, J. F., et al. (2013). Estimating absolute methylation levels at single-cpg resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome research*, 23(9):1541–1553.

Tatton-Brown, K., Seal, S., Ruark, E., Harmer, J., Ramsay, E., del Vecchio Duarte, S., Zachariou, A., Hanks, S., O'Brien, E., Aksglaede, L., et al. (2014). Mutations in the dna methyltransferase gene dnmt3a cause an overgrowth syndrome with intellectual disability. *Nature genetics*, 46(4):385–388.

Tian, Q., Zou, J., Tang, J., Fang, Y., Yu, Z., and Fan, S. (2019). Mrcnn: a deep learning model for regression of genome-wide dna methylation. *BMC genomics*, 20(2):1–10.

Tran, T.-A., Pham, D.-M., Ou, Y.-Y., et al. (2021). An extensive examination of discovering 5-methylcytosine sites in genome-wide dna promoters using machine learning based approaches. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(1):87–94.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Wang, H., Liu, H., Huang, T., Li, G., Zhang, L., and Sun, Y. (2022). Emdlp: Ensemble multiscale deep learning model for rna methylation site prediction. *Bmc Bioinformatics*, 23(1):221.

Xu, W., Yang, H., Liu, Y., Yang, Y., Wang, P., Kim, S.-H., Ito, S., Yang, C., Wang, P., Xiao, M.-T., et al. (2011). Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α-ketoglutarate-dependent dioxygenases. *Cancer cell*, 19(1):17–30.

Yu, Y., Si, X., Hu, C., and Zhang, J. (2019). A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270.

Yuan, F., Karatzoglou, A., Arapakis, I., Jose, J. M., and He, X. (2019). A simple convolutional generative network for next item recommendation. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 582–590.

Zhang, L., Lu, Q., and Chang, C. (2020a). Epigenetics in health and disease. *Epigenetics in allergy and autoimmunity*, pages 3–55.

Zhang, L., Xiao, X., and Xu, Z.-C. (2020b). ipromoter-5mc: a novel fusion decision predictor for the identification of 5-methylcytosine sites in genome-wide dna promoters. *Frontiers in Cell and Developmental Biology*, 8:614.

Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T., and Engelhardt, B. E. (2015). Predicting genome-wide dna methylation using methylation marks, genomic position, and dna regulatory elements. *Genome biology*, 16:1–20.

Zhou, X. (2020). *DNA methylation prediction model based on recurrent neural network and its fusion* *method*. PhD thesis, University of Electronic Science and Technology of China.