frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 25 June 2025
DOI 10.3389/fpls.2025.1626539

OPEN ACCESS

EDITED BY
Shanwen Sun,
Northeast Forestry University, China

REVIEWED BY
Lixin Cheng,
Jinan University, China
Fei Guo,
Central South University, China

*CORRESPONDENCE
Zhibin Lv
lvzhibin@pku.edu.cn

RECEIVED 11 May 2025
ACCEPTED 04 June 2025
PUBLISHED 25 June 2025

CITATION
Zhang Y, Chen H, Xiang S and Lv Z (2025) Identification of DNA N6-methyladenine modifications in the rice genome with a fine-tuned large language model. *Front. Plant Sci.* 16:1626539. doi: 10.3389/fpls.2025.1626539

Identification of DNA N6-methyladenine modifications in the rice genome with a fine-tuned large language model

Yichi Zhang, Hao Chen, Shicheng Xiang and Zhibin Lv*

College of Biomedical Engineering, Sichuan University, Chengdu, China

DNA N6-methyladenine (6mA) plays a significant role in various biological processes. In the rice genome, 6mA is involved in important processes such as growth and development, influencing gene expression. Therefore, identifying the 6mA locus in rice is crucial for understanding its complex gene expression regulatory system. Although several useful prediction models have been proposed, there is still room for improvement. To address this, we propose an architecture named iRice6mA-LMXGB that integrates a fine-tuned large language model to identify the 6mA locus in rice. Specifically, our method consists of two main components: (1) a BERT model for feature extraction and (2) an XGBoost module for 6mA classification. We utilize a pre-trained DNABERT-2 model to initialize the parameters of the BERT component. Through transfer learning, we fine-tune the model on the rice 6mA recognition task, converting raw DNA sequences into high-dimensional feature vectors. These features are then processed by an XGBoost algorithm to generate predictions. To further validate the effectiveness of our fine-tuning strategy, we employ UMAP(Uniform Manifold Approximation and Projection) visualization. Our approach achieves a validation accuracy of 0.9903 in a five-fold cross-validation setting and produces a receiver operating characteristic (ROC) curve with an area under the curve (AUC) of 0.9994. Compared to existing predictors trained on the same dataset, our method demonstrates superior performance. This study provides a powerful tool for advancing research in rice 6mA epigenetics.

KEYWORDS

rice genome, N6-methyladenine, large language model, BERT, UMAP visualization

1 Introduction

N6-methyladenine(6mA) is produced by methylation of the N6 position of adenine and has been found in bacteria, eukaryotes, and archaea (Zhang et al., 2015; O'Brown and Greer, 2016). Rice is one of the most important cereal crops in the world. Within the rice genome, 6mA serves as a critical epigenetic modification, regulating gene expression

through methylation at the N6 position of adenine (Lv et al., 2020; Chen et al., 2022; Jin et al., 2022). Studies have shown that 6mA in rice plays a vital role in many biological functions. For example, 6mA in rice is associated with stress response and helps rice to better adapt to adversity (Zhang et al., 2018; Ding et al., 2023). It is also associated with reproduction and regulates the growth and development of rice (Zhou et al., 2021; Yang et al., 2024). Zhou et al. discovered that 6mA is highly enriched in specific sequence motifs, conserved DNA sequence patterns that serve as recognition sites for epigenetic regulators. These motifs include AGG and GAGG, which are assumed to represent the binding elements of methyltransferase complexes or chromatin associated proteins. 6mA methylation preferentially occurred on these specific nucleotide motifs, indicating their functional significance in epigenetic regulation (Lee et al., 2018). And this methylation pattern is tightly linked to the drought stress response in rice (Zhou et al., 2018; Yang et al., 2024). In addition, 6mA can directly affect seed size and yield formation by regulating the expression of endosperm development-related genes (Zhou et al., 2021). In recent years, epigenetic breeding strategies based on CRISPR-6mA editing technology have provided new ideas to improve disease resistance and yield in rice by targeting modification of the 6mA locus (Romero and Gatica-Arias, 2019). However, traditional experimental methods such as SMRT-seq for detecting 6mA locus have the limitations of high cost and low throughput, and there is an urgent need to develop efficient computational prediction models to guide subsequent functional studies (Zhu et al., 2018; Wang L. et al., 2023; Chen et al., 2024; Liu et al., 2024; Shao et al., 2024; Xie H. et al., 2024; Zhou et al., 2024).

In recent years, machine and deep learning approaches have successfully addressed many challenges in identifying 6mA modifications in rice genomes (Sinha et al., 2023; Wang R. et al., 2023). In 2019, Chen et al. developed the first method for predicting DNA 6mA sites in rice, called i6mA-Pred, utilizing nucleotide chemical property (NCP) features and a support vector machine (SVM) as the classifier (Chen et al., 2019; Zou et al., 2022; Meher et al., 2024; Wang Y. et al., 2024). Subsequent research has seen the emergence of various single-classifier-based prediction methods, including MM-6mAPred (Pian et al., 2019), i6mA-DNCP (Park et al., 2020), iN6-methylat (Le, 2019), and iDNA6mA-rice (Lv et al., 2019). Moreover, ensemble learning models combining multiple classifiers, such as csDMA (Liu et al., 2019), SDM6A (Basith et al., 2019), 6mA-Finder (Xu et al., 2020), Meta-i6mA (Hasan et al., 2021), i6mA-VC (Xue et al., 2021), i6mA-Vote (Teng et al., 2022), and EpiSemble (Sinha et al., 2023), have been developed to enhance model performance and robustness. Deep learning techniques have evolved from traditional artificial neural network frameworks and have shown significant improvement in predictive power across multiple research domains. With the development of deep learning and its excellent performance, researchers began to apply it to the problem of DNA 6mA site prediction. In 2019, Yu et al. developed a prediction model called SNNRice6mA (Yu and Dai, 2019) based on convolutional neural networks (CNNs) through single-nucleotide one-hot coding, obtaining an accuracy of 0.920. Another group of researchers, Lv et al., proposed a convolutional neural network iRicem6A-CNN (Lv et al., 2021) based on a dinucleotide one-hot encoder in 2020, achieving an accuracy of 0.938 for 5-fold cross-validation. However, it is worth noting that CNNs are limited in focusing on only part of the information. Deep6mA (Li Z. et al., 2021), which consists of a convolutional neural network (CNN) and a bidirectional LSTM (BLSTM) module to solve the long-distance nucleotide association problem by learning contextual dependencies of the sequences, was proposed by Li et al. in 2021 and achieved a 5-fold cross-validation accuracy of 0.940.

Over the last few years, large-scale language modeling (LLM) has progressed tremendously (Li H. et al., 2021; Xie X. et al., 2024; Chen et al., 2025). The well-known model, ChatGPT, is a fine-tuned version of the base GPT-3 model. By learning contextual text in a self-supervised manner, it can both understand and generate human language (Devlin et al., 2019; Wang G. et al., 2024). DNA sequences exhibit similarities to natural language. Nucleotides, the building blocks of nucleic acids, serve as "words" within biological systems' "languages". LLMs can be adapted for the analysis of biological sequence data by leveraging the structure of DNA and protein sequences as analogous to natural language texts (Jumper et al., 2021; Rives et al., 2021; Wei et al., 2021; Li T. et al., 2024; Li Y. et al., 2024; Qiao et al., 2024; Lai et al., 2025; Xie et al., 2025). There have been many breakthroughs in LLMs for applications in biology, such as AlphaFold2, a protein prediction model with very high accuracy (Jumper et al., 2021), the Geneformer model trained on data from about 10 million human single-cell RNA sequences (Zou et al., 2019; Theodoris et al., 2023), and DNABERT, a transformer-based DNA pre-training model (Ji et al., 2021). While LLMs demonstrate potential for identifying patterns and correlations in noisy biological datasets (Lam et al., 2024; Soylu and Sefer, 2024; Xie X. et al., 2024; Liu et al., 2025), they have yet to gain acceptance within plant science research. To date, LLMs have not been employed in the study of 6mA locus prediction in rice.

In this study, we develop a large language model-based transfer learning model called iRice6mA-LMXGB. it consists of a pre-trained DNABERT2 model and an XGBoost model. It contains a unique fine-tuning architecture that relies exclusively on DNA sequence data to distinguish 6mA sequences from non-6mA sequences. Experimental results demonstrate the model's outstanding performance, achieving a validation accuracy of 0.9903 through 5-fold cross-validation. Compared to all previous methods tested on standard datasets, iRice6mA-LMXGB significantly outperforms them, suggesting that this novel approach has the potential to transform biological sequence modeling.

2 Materials and methods

2.1 Benchmark dataset

In this study, we utilized the rice dataset constructed by Lv et al. (2020) for model training and evaluation using 5-fold cross-validation. To ensure the high quality of the data, sequences with greater than 80% similarity were removed via the CD-HIT program

(Li and Godzik, 2006). The dataset is made of 154,000 sequences with 6mA sites and 154,000 sequences without 6mA sites. This is a widely adopted and balanced rice dataset. During model training, unbalanced datasets may lead to unreliable results. The majority class samples are dominant and the model will favor the majority class during training, thus ignoring the minority class. This may result in the model having high accuracy for the majority class but low recognition for the minority class during prediction. For ease of reference, we denote it as "rice-Lv" throughout this study. Both positive and negative sequences in the rice-Lv dataset are 41 base pairs in length. Positive sequences represent 6mA modifications at their centers, while negative sequences lack such modifications at theirs. By employing this well-established dataset, we enable a fair comparison between our method and those previously reported.

2.2 Architecture of iRice6mA-LMXGB

The architecture of iRice6A-LMXGB is presented in Figure 1, comprising two main components: the pre-trained DNABERT-2 module and the XGBoost module. DNABERT-2 is a pre-trained BERT model specifically designed for encoding DNA sequences. It can efficiently identify complex long-range dependencies in these sequences (Zhou et al., 2023). And this module will undergo further fine-tuning in this study. XGBoost's superior performance, particularly in terms of speed and accuracy when processing large-scale datasets, enables its extensive use in solving classification problems (Chen and Guestrin, 2016; Yang et al., 2021). It utilizes the feature vectors output from the DNABERT-2 model to generate final prediction results. A detailed explanation of the model follows.

2.2.1 DNABERT-2

DNABERT-2 is an iterative version of DNABERT. DNABERT is the first BERT-based DNA language model (Ji et al., 2021). Rigorously trained on a comprehensive genomic dataset encompassing the entire human genome, DNABERT offers a linguistic perspective for genomic analysis. While widely adopted, the initial version of DNABERT exhibited notable technical limitations. Specifically, DNABERT faced two critical challenges: first, its training data is limited to a single-species genome, which makes it difficult for the model to capture sequence-conserving patterns and diversity features across species; second, the k-mer sequence partitioning mechanism it employs not only triggers the hidden danger of data leakage during the training process, but also significantly increases the computational complexity (Moeckel et al., 2024). Such limitations underscore the pressing need for innovation and improvement in DNA-based language modeling research. To address these challenges, DNABERT-2 introduced significant improvements in both areas. First, it breaks through species boundaries and employs cross-species genomic datasets for pre-training, significantly enhancing the model's ability to recognize evolutionarily conserved regions and species specificity. Second, at the data processing stage, DNABERT-2 employs byte-pair encoding (BPE), a novel tokenization method that replaces traditional k-mer partitioning. This is a data compression algorithm widely used in large-scale language models (Sennrich et al., 2015), which effectively solves the risk of data leakage and improves computational efficiency, successfully overcoming the limitations of k-mer tokenization. As demonstrated by Zhihan et al.'s comparative analysis, compared to conventional 6-mer tokenization methods, the byte-pair encoding (BPE) implementation exhibits superior sequence compression efficiency, reducing the tokenized sequence length by a factor of 5. The dramatic reduction in dimensionality directly improves the computational efficiency of processing genome sequences (Zhou et al., 2023).

The BERT model consists of two independent components: the module responsible for preprocessing BERT input and the pre-training BERT module. In the BERT input preprocessing module, DNABERT-2 utilizes BPE to tokenize DNA sequences. Byte Pair Encoding (BPE) is a subword tokenization algorithm commonly employed in NLP Natural Language Processing) tasks. Its key mechanism lies in iteratively merging character pairs of the highest frequency to construct a vocabulary of subwords. During tokenization, DNABERT-2 appends a [CLS] token at the sequence start and a [SEP] token at the end. Then, each token is put into an embedding module and converted into a vector. The DNABERT-2 model uses the ALiBi(Attention with Linear Biases) (Press et al., 2021) approach, which does not add positional embeddings to the input, but rather adds a non-learned embedding in every Attention computation to add a non-learning bias and a fixed set of statics to combine the location information with the Attention score. DNABERT-2 employs a transformer encoder architecture as the backbone of its pre-trained BERT module. The feature matrix is constructed by cascading encoders layer by layer across the network's layers (L). Each encoder comprises three components: multi-head self-attention units, position-wise feed-forward neural networks, and normalization layers. Within the i-th encoder stage, the multi-head self-attention mechanism operates as follows.

$$\text{Multihead}(X^i) = \text{Concat}(head_1, head_2, \ldots, head_n)W^{O,i}$$

For the i-th encoder, the input matrix X^i is handled through n self-attentive heads for processing. The outputs of these heads are then transformed by the output transformation matrix $W^{o,i}$, which is computed in detail for each $head^i$ as follows.

$$\text{head}^i = \text{softmax}\left(\frac{W^{Q,i}X^i(W^{K,i}X^i)^T}{\sqrt{d_k}}\right)W^{V,i}X^i$$

$W^{Q,i}$, $W^{K,i}$ and $W^{V,i}$ serve as the transformation matrices for the query, key, and value components of each head, respectively. d_k denotes the dimension of the matrix.

Specifically, after computing $MultiHead(X^i)$ in the multi-head attention mechanism, this resultant output is added to the residual connection of the original input X^i for normalization. The computation proceeds according to the formula below.

$$Y^i = \text{LayerNorm}(\text{MultiHead}(X^i) + X^i)$$

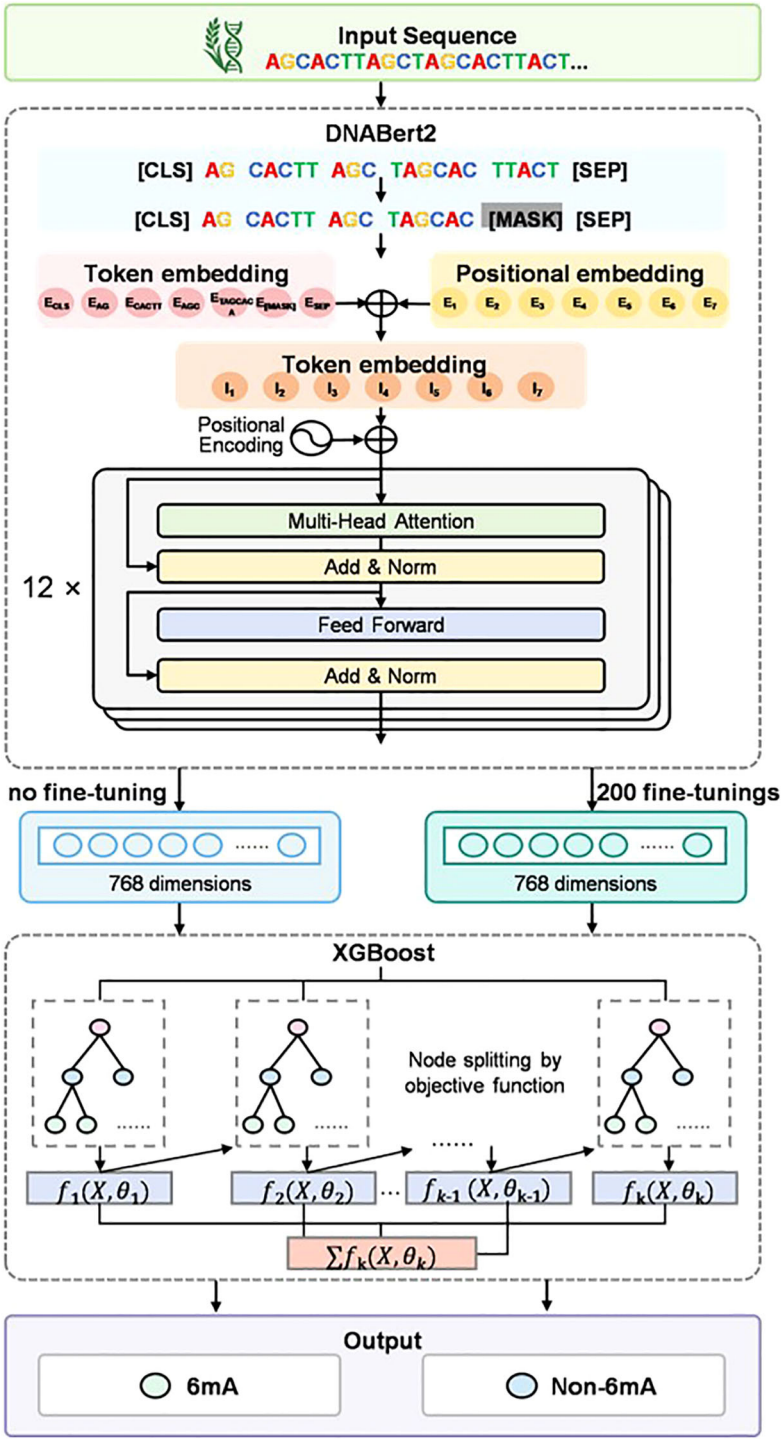

FIGURE 1
The proposed modeling framework, iRice6mA-LMXGB.

After normalization, the processed data is passed through a feed-forward neural network using the following formula:

$$FFN(Y^i) = \max(0, Y^i W_1 + b_1)W_2 + b_2$$

W_1, W_2, b_1 and b_2 are the trainable weight parameters within the feed-forward layer.

The output of the i-th encoder is achieved through normalization of the residual connection between Y^i and $FFN(Y^i)$. Below is the corresponding formula.

$$X^{i+1} = \text{LayerNorm}(Y^i + \text{FFN}(Y^i))$$

Finally, the output of the DNABERT-2 can be obtained by cascading the L encoders as follows.

$$X_1 = X^{i+1} \in R^{d \times N}$$

where d denotes the dimension of the word vector and N represents the total number of tokens.

DNABERT-2 follows the BERT model architecture, defined by three key parameters: L = 12, H = 768, and A = 12. The parameter L specifies the number of transformer layers (totaling 12). The parameter H determines the hidden layer size, with each token represented as a 768-dimensional vector. The parameter A specifies the number of attention heads (totaling 12). In this study, we use the full fine-tuning (FFT) (Church et al., 2021) method, which treats rice DNA sequences as "sentences in natural language" and inputs them into the DNABERT-2 module to adjust and update all the parameters. Finally, we use the BERT model to convert them into fixed-length feature vectors to obtain the original feature matrix before fine-tuning and the feature matrix after 200 cycles of updating.

2.2.2 XGBOOST

The XGBoost classifier is a gradient boosting method that integrates regression trees (Basith et al., 2019). The objective function of the model is $obj(\theta) = L(\theta) + \Omega(\theta)$, $L(\theta)$ is the training loss function with the expression:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i)$$

Where $l(y_i, \hat{y}_i)$ represents the training loss function for each sample. y_i represents the true value of the i-th sample. $\hat{y}_i$ represents the estimated value of the i-th sample.

Then the estimated value of the i-th sample is expressed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F$$

K is the number of integrated trees, and F denotes the space of all possible decision trees. f_k is a specific categorical regression tree (CART).$\Omega(f)$ is the tree structure complexity function, and its specific form is:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{i=1}^{T} w_i^2$$

The parameter γ restricts the number of leaf nodes T of the tree to control the complexity of the model. And the parameter λ constrains the sum of the weights w_i^2 of each leaf node to suppress overfitting. The objective function is continuously optimized by adjusting the parameters for the optimal result. In this way, the XGBoost classifier finally outputs the prediction results of the rice sequence about 6mA by receiving the extracted feature vectors from DNABERT-2.

2.3 Evaluation metrics and methods

In this study, we validate our approach using a traditional 5-fold cross-validation method and compare it to previous studies based on the benchmark dataset rice-Lv.

we will combine five metrics, including accuracy (ACC), sensitivity (Sn), specificity (Sp), Matthew's correlation coefficient (MCC), and area under the curve (AUC), to comprehensively evaluate the prediction performance of our model (Zou et al., 2023; Zulfiqar et al., 2023; Guo et al., 2024; Huang et al., 2024; Zhu et al., 2024).

ACC indicates the overall correctness of the model prediction and is a basic benchmark used to evaluate the model performance, which can be expressed as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

The sensitivity Sn, also known as the true positive rate (TPR), is expressed as:

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

The specificity Sp, also known as the true negative rate (TNR), is expressed as:

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

MCC is a composite metric that assesses the overall quality of classification model predictions by examining the performance of the classification model in each of the four quadrants of the confusion matrix. The superior score reflects the balanced excellence between true positives (TP), true negatives (TN), false negatives (FN) and false positives (FP). It can be defined as:

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

The last performance metric we use is AUC, defined as the value of the area under the subject's operating characteristic curve. AUC is also an important measure of the performance of a dichotomous model. The larger the value of AUC, the better the model performs. AUC is a floating-point number between 0 and 1. 1 indicates that the model predicts perfectly, whereas 0.5 indicates that the model is similar to a random prediction (Zhang et al., 2025).

3 Results and discussion

3.1 Model performance analysis

In this study, we developed three models. For the first model, we directly used the pre-trained DNABERT-2 to extract 768-dimensional features from rice DNA and fed them into an XGBoost classifier for prediction tasks. The XGBoost classifier shows unique advantages in genomics data classification tasks, mainly due to its ability to efficiently handle high-dimensional sparse data and its built-in regularization mechanism. Our dataset, with more than 300,000 samples, is characterized by high feature dimensionality, and XGBoost is able to efficiently capture nonlinear interaction effects through the gradient boosting

framework combined with second-order derivative optimization. Its regularization term can in turn suppress overfitting and enhance model generalization (Chen and Guestrin, 2016). Cross-validation results showed ACC=0.6259, Sn=0.6207, Sp=0.6312, MCC=0.2519, and auROC=0.6728 for this configuration. For the second model, we loaded the rice-Lv dataset into the DNABERT-2 module and conducted 200 iteration loops to develop a fine-tuned version of the model. The 5-fold cross-validation scores were ACC=0.9903, Sn=0.9898, Sp=0.9907, MCC=0.9805, auROC=0.9994 which are 58.22%, 59.47%, 56.96%, 289.24%, and 48.54%, respectively higher than those of the non-fine-tuned model. For the third model, we utilized LightGBM's built-in function to assess and prioritize feature importance using features extracted from the fine-tuned DNABERT-2 model (Ke et al., 2017). The feature ranking principle of LightGBM is based on the Gradient Boosting Decision Tree (GBDT) framework, which evaluates feature importance by quantifying the contribution of features in the process of constructing the decision tree (Ke et al., 2017). Following this, we selected the top 300 features for modeling with XGBoost. The 5-fold cross-validation yielded ACC=0.9899, Sn=0.9890, Sp=0.9908, MCC=0.9799, and auROC=0.9994. As shown in Figure 2, our cross-validation results indicate that: (1) Fine-tuned models outperformed non-fine-tuned counterparts significantly. (2) However, applying feature selection after fine-tuning caused minor performance degradation compared to models without feature selection, not much difference overall. These findings demonstrate the effectiveness of our fine-tuning strategy. The pre-training model is usually trained on multi species datasets, and may not be able to capture the 6mA distribution pattern unique to rice. Through the fine-tuning strategy, the model parameters are recalibrated, which can give priority to the local features in the rice genome, and the sensitivity of the model to the sequence context of rice 6mA is improved. Additionally, while XGBoost's tree-based architecture excels at managing high-dimensional data through regularization techniques, our results suggest that applying LightGBM-based feature selection after fine-tuning may slightly reduce model performance due to fewer feature interactions. We selected the second model with the best performance, performing fine-tuning for 200 iterations without feature selection, to name iRice6mA-LMXGB.

3.2 UMAP dimensionality reduction visualization

In order to perform an in-depth analysis of the interpretability of the iRice6mA-LMXGB model after integrating DNABERT-2 with XGBoost, we used the UMAP (Uniform Manifold Approximation and Projection) technique. This is a nonlinear dimensionality reduction and visualization algorithm for large-scale datasets. Umap assumes that the data is distributed on a low dimensional manifold. Firstly, the probability weight is defined in the high dimensional space using the neighborhood graph to reflect the similarity between points. Then the cross entropy loss function is used to optimize the embedding in the low dimensional space to align the low dimensional similarity with the high dimensional structure. Based on graph theory and flow learning methods, it is assumed that the available data samples are uniformly distributed in the topological space and can be approximated and mapped from these finite data samples to a lower-dimensional space for visualization and analysis (McInnes and Healy, 2018).

To be more specific, we will visualize the distribution of 6mA and non-6mA by projecting each feature vector onto a 2D view using the UMAP technique. Figure 3 shows the arrangement of 6mA and non-6mA samples in 2D space before and after fine-tuning, and the decision boundary drawn in black by the XGBoost algorithm. Blue markers denote non-6mA samples, and orange markers denote 6mA samples. The first subplot represents the UMAP results of the original features without fine-tuning, which can be interpreted as all the sample points not showing any representative clustering. In Figure 3A, poor

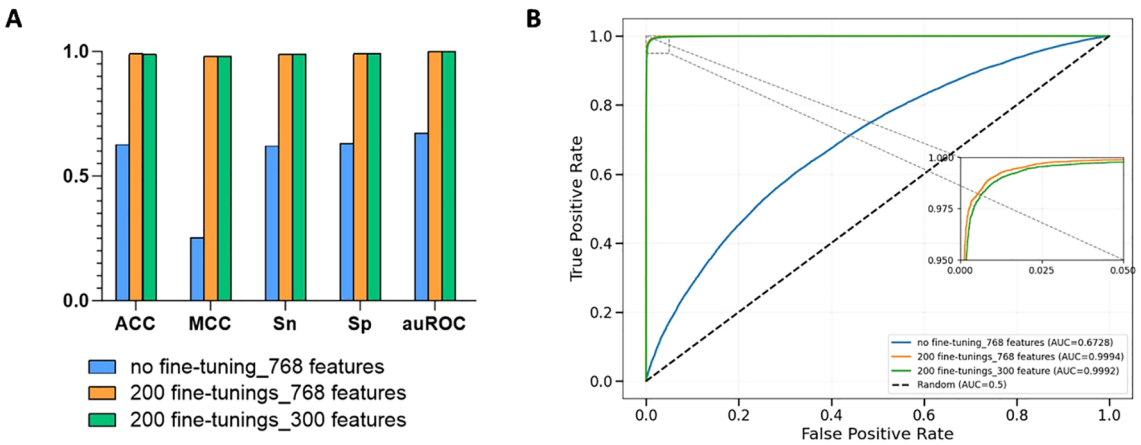

FIGURE 2
(A) Comparison of model performance with or without fine-tuning and with or without feature selection; **(B)** Average ROC curves for five-fold cross-validation of the three models. Where no fine-tuning_768 features denotes the model with no fine-tuning, 200 fine-tunings_768 features denotes the model with two hundred fine-tunings without feature selection, and 200 fine-tunings_300 features denotes the model that was fine-tuned 200 times and ranked for feature importance and the top 300 features are selected after the feature importance ranking.

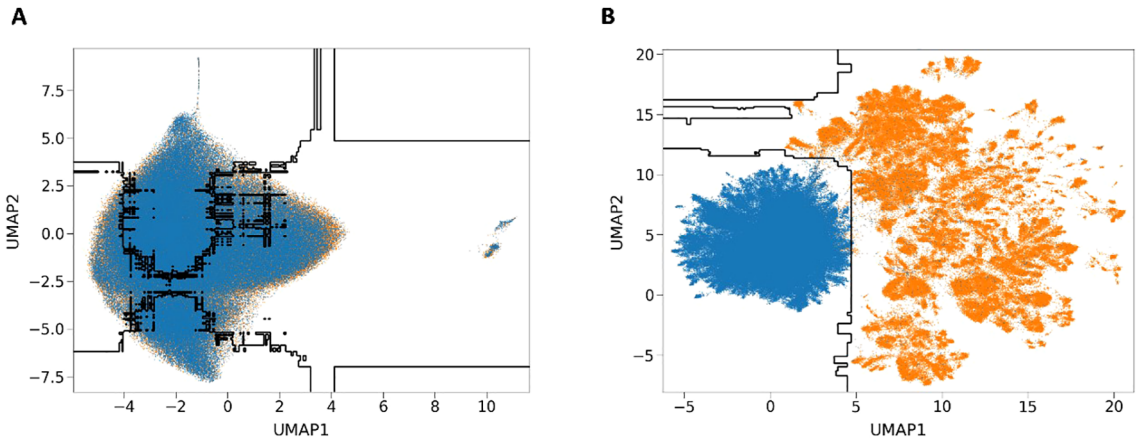

FIGURE 3
UMAP dimensionality reduction visualization. **(A)** UMAP results of the original features of the unfine-tuned model. **(B)** UMAP results of features learned by the iRice6mA-LMXGB model.

separation indicates significant feature overlap between the 6mA sample points and the non-6mA sample points (Figure 3A), suggesting a high degree of overlap in their distributions. The second subfigure shows the results of projecting the high-dimensional feature space learned from the iRice6mA-LMXGB model into a 2D view, which shows much improved clustering, indicating a significant increase in separation and a decrease in overlap in the feature space (Figure 3B), resulting in improved performance. In summary, our approach allows for better learning of model decision boundaries. Through this visualization technique, we can more intuitively understand the impact of features on model predictions, further deepening our exploration of model interpretability.

3.3 Comparison of the proposed model with existing models

To better evaluate the performance of our model, we compare it with the following state-of-the-art methods, including MM-6mAPred (Pian et al., 2019), iDNA6mA-Rice (Lv et al., 2019), SNNRice6mA (Yu and Dai, 2019), iRicem6A-CNN (Lv et al., 2021), ENet-6mA (Abbas et al., 2022), Deep6mA (Li Z. et al., 2021) and SpineNet-6mA (Abbas et al., 2020). Our model is evaluated using the same five-fold cross-validation protocol on the same dataset as previous studies, employing the identical metrics: ACC, MCC, Sn, Sp, and AUC. As shown in Table 1, our iRice6mA-LMXGB model outperforms all previous predictors across all metrics and demonstrates more stable performance with less fluctuation in ACC, MCC, Sn, Sp, and AUC values. In ACC, MCC, Sn, and AUROC metrics, our model improves over the previous best predictor SpineNet-6mA by 5%, 11.42%, 3.42%, 6.62%, and 1.98%, respectively. Furthermore, it outperforms the previous best model, ENet-6mA, by 6.08% in Sp metric. To facilitate visualization of the comparison results, we created a box-and-whisker plot, as illustrated in Figure 4. To sum up, our iRice6mA-LMXGB model demonstrates superior performance compared to both machine learning-based and CNN/LSTM-based deep learning models for 6mA prediction in rice, showcasing its robustness as a predictive tool.

4 Conclusions

In this article, we develop a novel computational model called iRice6mA-LMXGB that combines fine-tuned large language modeling to efficiently distinguish and identify 6mA and non-6mA loci in the rice genome. We utilized the large language model, DNABERT-2, to represent the DNA sequence as a continuous word vector, thus effectively capturing the DNA sequence features. Subsequently, we applied the robust machine learning method XGBoost to make accurate predictions based on the extracted features. We compare and analyze the performance of iRice6mA-LMXGB with other predictors, and the results show that iRice6mA-LMXGB obtains the best performance compared to previous models. Our model outperforms all existing models on ACC, SN, SP, MCC, and AUC (5-fold cross-validation: ACC = 0.9903, MCC = 0.9805, Sn = 0.9898, Sp

TABLE 1 5-fold cross-validation results of iRice6mA-LMXGB with several previous methods on the rice-Lv dataset.

Method	ACC	MCC	Sn	Sp	AUROC
MM-6mAPred	0.9149	0.8300	0.9347	0.8951	0.9600
iDNA6mA-Rice	0.9170	0.8350	0.9300	0.9050	0.9640
SNNRice6mA	0.9204	0.8400	0.9433	0.8975	0.9700
iRicem6A-CNN	0.9382	0.8770	0.9434	0.9331	0.9790
ENet-6mA	0.9437	0.8700	0.9467	0.9339	0.9800
Deep6mA	0.9401	0.8800	0.9506	0.9296	0.9800
SpineNet-6mA	0.9431	0.8800	0.9571	0.9292	0.9800
iRice6mA-LMXGB (ours)	**0.9903**	**0.9805**	**0.9898**	**0.9907**	**0.9994**

Bold values indicate that the model proposed in this study achieves optimal results in each of the assessment metrics.

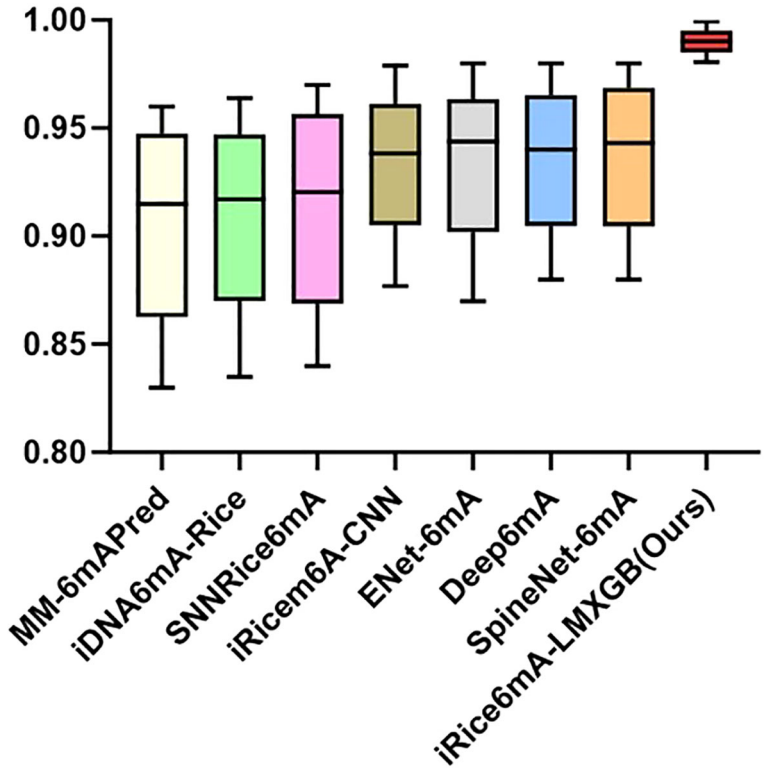

FIGURE 4
Comparison of the proposed model with other existing models on the rice-Lv dataset.

= 0.9907, and auROC = 0.9994), suggesting that the iRice6mA-LMXGB is a powerful and robust predictor that can help researchers to identify and analyze the 6mA locus in the rice genome more effectively, thus providing a deeper understanding of the complex mechanisms of gene regulation and advancing the field of life sciences. It is demonstrated through UMAP visualization that the fine-tuning strategy for large language models significantly enhances the model's feature extraction ability. This raises the possibility that large language models can be fine-tuned for various purposes and deployed for plant-specific domains to solve biological problems. Moving ahead, we plan to expand our dataset and perform model optimization to enhance the generalizability of our model for broader applications.

Data availability statement

The raw sequence data used in the study were obtained from the following URL: http://lin-group.cn/server/iDNA6mA-Rice.

Author contributions

YZ: Formal analysis, Investigation, Visualization, Writing – original draft. HC: Investigation, Writing – review & editing. SX: Investigation, Writing – review & editing. ZL: Methodology, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work supports by the National Natural Science Foundation of China (No.62371318, No.62001090) and 2024 Foundation Cultivation Research Basic Research Cultivation Special Funding (No. 20826041H4211).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

Abbas, Z., Tayara, H., and Chong, K. T. (2020). SpineNet-6mA: A novel deep learning tool for predicting DNA N6-methyladenine sites in genomes. *IEEE Access* 8, 201450–201457. doi: 10.1109/Access.6287639

Abbas, Z., Tayara, H., and Chong, K. T. (2022). ENet-6mA: identification of 6mA modification sites in plant genomes using elasticNet and neural networks. *Int. J. Mol. Sci.* 23, 8314. doi: 10.3390/ijms23158314

Basith, S., Manavalan, B., Shin, T. H., and Lee, G. (2019). SDM6A: A web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Mol. Ther. Nucleic Acids* 18, 131–141. doi: 10.1016/j.omtn.2019.08.011

Chen, T., and Guestrin, C. (2016). "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA. 785–794 (Association for Computing Machinery).

Chen, B., Guo, Y., Zhang, X., Wang, L., Cao, L., Zhang, T., et al. (2022). Climate-responsive DNA methylation is involved in the biosynthesis of lignin in birch. *Front. Plant Sci.* 13, 1090967. doi: 10.3389/fpls.2022.1090967

Chen, L., Liu, G., and Zhang, T. (2024). Integrating machine learning and genome editing for crop improvement. *aBIOTECH* 5, 262–277. doi: 10.1007/s42994-023-00133-5

Chen, W., Lv, H., Nie, F., and Lin, H. (2019). i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics* 35, 2796–2800. doi: 10.1093/bioinformatics/btz015

Chen, S., Yan, K., Li, X., and Liu, B. (2025). Protein language pragmatic analysis and progressive transfer learning for profiling peptide–protein interactions. *IEEE Trans. on Neural Networks and Learn. Syst* 2025, 1–15. doi: 10.1109/TNNLS.2025.3540291

Church, K. W., Chen, Z., and Ma, Y. (2021). Emerging trends: A gentle introduction to fine-tuning. *Nat. Lang. Eng.* 27(6), 763–778. doi: 10.1017/S1351324921000322

Devlin, J., Chang, M.–W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, volume 1 (Long and Short Papers): 4171–4186. doi: 10.18653/v1/N19-1423

Ding, K., Sun, S., Luo, Y., Long, C., Zhai, J., Zhai, Y., et al. (2023). PlantCADB: A comprehensive plant chromatin accessibility database. *Genom Proteomics Bioinf.* 21, 311–323. doi: 10.1016/j.gpb.2022.10.005

Guo, X., Huang, Z., Ju, F., Zhao, C., and Yu, L. (2024). Highly accurate estimation of cell type abundance in bulk tissues based on single-cell reference and domain adaptive matching. *Adv Sci.* 11, 2306329. doi: 10.1002/advs.202306329

Hasan, M. M., Basith, S., Khatun, M. S., Lee, G., Manavalan, B., and Kurata, H. (2021). Meta-i6mA: an interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. *Brief Bioinform.* 22, bbaa202. doi: 10.1093/bib/bbaa202

Huang, Z., Guo, X., Qin, J., Gao, L., Ju, F., Zhao, C., et al. (2024). Accurate RNA velocity estimation based on multibatch network reveals complex lineage in batch scRNA-seq data. *BMC Biol.* 22, 290. doi: 10.1186/s12915-024-02085-8

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* 37, 2112–2120. doi: 10.1093/bioinformatics/btab083

Jin, J., Yu, Y., Wang, R., Zeng, X., Pang, C., Jiang, Y., et al. (2022). iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations. *Genome Biol.* 23, 1–23. doi: 10.1186/s13059-022-02780-1

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi: 10.1038/s41586-021-03819-2

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.–Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process Syst.* 30, 3146–3154. doi: 10.5555/3294996.3295074

Lai, L., Liu, Y., Song, B., Li, K., and Zeng, X. (2025). Deep generative models for therapeutic peptide discovery: A comprehensive review. *ACM Comput. Surv* 57, 1–29. doi: 10.1145/3714455

Lam, H. Y. I., Ong, X. E., and Mutwil, M. (2024). Large language models in plant biology. *Trends Plant Sci.* 29, 1145–1155. doi: 10.1016/j.tplants.2024.04.013

Le, N. Q. K. (2019). 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* 294, 1173–1182. doi: 10.1007/s00438-019-01570-y

Lee, N. K., Li, X., and Wang, D. (2018). A comprehensive survey on genetic algorithms for DNA motif prediction. *Inf. Sci.* 466, 25–43. doi: 10.1016/j.ins.2018.07.004

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Li, Z., Jiang, H., Kong, L., Chen, Y., Lang, K., Fan, X., et al. (2021). Deep6mA: A deep learning framework for exploring similar patterns in DNA N6-methyladenine sites across different species. *PloS Comput. Biol.* 17, e1008767. doi: 10.1371/journal.pcbi.1008767

Li, H., Pang, Y., and Liu, B. (2021). BioSeq-BLM: a platform for analyzing DNA, RNA, and protein sequences based on biological language models. *Nucleic Acids Res.* 49, e129. doi: 10.1093/nar/gkab829

Li, T., Ren, X., Luo, X., Wang, Z., Li, Z., Luo, X., et al. (2024). A foundation model identifies broad-spectrum antimicrobial peptides against drug-resistant bacterial infection. *Nat. Commun.* 15, 7538. doi: 10.1038/s41467-024-51933-2

Li, Y., Wei, X., Yang, Q., Xiong, A., Li, X., Zou, Q., et al. (2024). msBERT-Promoter: a multi-scale ensemble predictor based on BERT pre-trained model for the two-stage prediction of DNA promoters and their strengths. *BMC Biol.* 22, 126. doi: 10.1186/s12915-024-01923-z

Liu, G., Chen, L., Wu, Y., Han, Y., Bao, Y., and Zhang, T. (2025). PDLLMs: A group of tailored DNA large language models for analyzing plant genomes. *Mol. Plant* 18, 175–178. doi: 10.1016/j.molp.2024.12.006

Liu, Z., Dong, W., Jiang, W., and He, Z. (2019). csDMA: an improved bioinformatics tool for identifying DNA 6 mA modifications via Chou's 5-step rule. *Sci. Rep.* 9, 13109. doi: 10.1038/s41598-019-49430-4

Liu, Y., Shen, X., Gong, Y., Liu, Y., Song, B., and Zeng, X. (2024). Sequence Alignment/Map format: a comprehensive review of approaches and applications. *Briefings Bioinf.* 24, bbad320. doi: 10.1093/bib/bbad320

Lv, H., Dao, F. Y., Guan, Z. X., Zhang, D., Tan, J. X., Zhang, Y., et al. (2019). iDNA6mA-rice: A computational tool for detecting N6-methyladenine sites in rice. *Front. Genet.* 10, 793. doi: 10.3389/fgene.2019.00793

Lv, Z., Ding, H., Wang, L., and Zou, Q. (2021). A convolutional neural network using dinucleotide one-hot encoder for identifying DNA N6-methyladenine sites in the rice genome. *Neurocomputing* 422, 214–221. doi: 10.1016/j.neucom.2020.09.056

Lv, H., Zhang, Z. M., Li, S. H., Tan, J. X., Chen, W., and Lin, H. (2020). Evaluation of different computational methods on 5-methylcytosine sites identification. *Brief Bioinform.* 21, 982–995. doi: 10.1093/bib/bbz048

McInnes, L., and Healy, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv (USA)*, abs/1802.03426. doi: 10.48550/arXiv.1802.03426

Meher, P. K., Hati, S., Sahu, T. K., Pradhan, U., Gupta, A., and Rath, S. N. (2024). SVM-root: identification of root-associated proteins in plants by employing the support vector machine with sequence-derived features. *Curr. Bioinf.* 19, 91–102. doi: 10.2174/1574893618666230417104543

Moeckel, C., Mareboina, M., Konnaris, M. A., Chan, C. S. Y., Mouratidis, I., Montgomery, A., et al. (2024). A survey of k-mer methods and applications in bioinformatics. *Comput. Struct. Biotechnol. J.* 23, 2289–2303. doi: 10.1016/j.csbj.2024.05.025

O'Brown, Z. K., and Greer, E. L. (2016). N6-methyladenine: A conserved and dynamic DNA mark. *Adv. Exp. Med. Biol.* 945, 213–246. doi: 10.1007/978-3-319-43624-1_10

Park, S., Wahab, A., Nazari, I., Ryu, J. H., and Chong, K. T. (2020). i6mA-DNC: Prediction of DNA N6-Methyladenosine sites in rice genome based on dinucleotide representation using deep learning. *Chemom Intell Lab. Syst.* 204, 104102. doi: 10.1016/j.chemolab.2020.104102

Pian, C., Zhang, G., Li, F., and Fan, X. (2019). MM-6mAPred: identifying DNA N6-methyladenine sites based on Markov model. *Bioinformatics* 36, 388–392. doi: 10.1093/bioinformatics/btz556

Press, O., Smith, N. A., and Lewis, M. (2021). Train short, test long: attention with linear biases enables input length extrapolation. *arXiv (USA)*, abs/2108.12409. doi: 10.48550/arXiv.2108.12409

Qiao, B., Wang, S., Hou, M., Chen, H., Zhou, Z., Xie, X., et al. (2024). Identifying nucleotide-binding leucine-rich repeat receptor and pathogen effector pairing using

transfer-learning and bilinear attention network. *Bioinformatics* 40, btae581. doi: 10.1093/bioinformatics/btae581

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. United States America* 118, e2016239118. doi: 10.1073/pnas.2016239118

Romero, F. M., and Gatica-Arias, A. (2019). CRISPR/cas9: development and application in rice breeding. *Rice Sci.* 26, 265–281. doi: 10.1016/j.rsci.2019.08.001

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany. Association for Computational Linguistics. (Volume 1: Long Papers): 1715–1725, doi: 10.18653/v1/P16-1162

Shao, M., Tian, M., Chen, K., Jiang, H., Zhang, S., Li, Z., et al. (2024). Leveraging random effects in cistrome-wide association studies for decoding the genetic determinants of prostate cancer. *Adv Sci.* 11, 2400815. doi: 10.1002/advs.202400815

Sinha, D., Dasmandal, T., Yeasin, M., Mishra, D. C., Rai, A., and Archak, S. (2023). EpiSemble: A novel ensemble-based machine-learning framework for prediction of DNA N6-methyladenine sites using hybrid features selection approach for crops. *Curr. Bioinf.* 18, 587–597. doi: 10.2174/1574893618666230316151648

Soylu, N. N., and Sefer, E. (2024). DeepPTM: protein post-translational modification prediction from protein sequences by combining deep protein language model with vision transformers. *Curr. Bioinf.* 19, 810–824. doi: 10.2174/0115748936283134240109054157

Teng, Z., Zhao, Z., Li, Y., Tian, Z., Guo, M., Lu, Q., et al. (2022). i6mA-vote: cross-species identification of DNA N6-methyladenine sites in plant genomes based on ensemble learning with voting. *Front. Plant Sci.* 13, 845835. doi: 10.3389/fpls.2022.845835

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., et al. (2023). Transfer learning enables predictions in network biology. *Nature* 618, 616–624. doi: 10.1038/s41586-023-06139-9

Wang, L., Ding, Y., Tiwari, P., Xu, J., Lu, W., Muhammad, K., et al. (2023). A deep multiple kernel learning-based higher-order fuzzy inference system for identifying DNA N4-methylcytosine sites. *Inf. Sci.* 630, 40–52. doi: 10.1016/j.ins.2023.01.149

Wang, R., Jiang, Y., Jin, J., Yin, C., Yu, H., Wang, F., et al. (2023). DeepBIO: an automated and interpretable deep-learning platform for high-throughput biological sequence prediction, functional annotation and visualization analysis. *Nucleic Acids Res.* 51, 3017–3029. doi: 10.1093/nar/gkad055

Wang, G., Lou, X., Guo, F., Kwok, D., and Cao, C. (2024). EHR-HGCN: an enhanced hybrid approach for text classification using heterogeneous graph convolutional networks in electronic health records. *IEEE J. Biomed. Health Inf.* 28, 1668–1679. doi: 10.1109/JBHI.2023.3346210

Wang, Y., Zhai, Y., Ding, Y., and Zou, Q. (2024). SBSM-Pro: support bio-sequence machine for proteins. *Sci. China-Inf Sci.* 67, 212106. doi: 10.1007/s11432-024-4171-9

Wei, L., He, W., Malik, A., Su, R., Cui, L., and Manavalan, B. (2021). Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework. *Brief Bioinform.* 22, bbaa275. doi: 10.1093/bib/bbaa275

Xie, H., Ding, Y., Qian, Y., Tiwari, P., and Guo, F. (2024). Structured Sparse Regularization based Random Vector Functional Link Networks for DNA N4-methylcytosine sites prediction. *Expert Syst. Appl.* 235, 121157. doi: 10.1016/j.eswa.2023.121157

Xie, X., Gui, L., Qiao, B., Wang, G., Huang, S., Zhao, Y., et al. (2024). Deep learning in template-free de novo biosynthetic pathway design of natural products. *Brief Bioinform.* 25, bbae495. doi: 10.1093/bib/bbae495

Xie, H., Wang, L., Qian, Y., Ding, Y., and Guo, F. (2025). Methyl-GP: accurate generic DNA methylation prediction based on a language model and representation learning. *Nucleic Acids Res.* 53, gkaf223. doi: 10.1093/nar/gkaf223

Xu, H., Hu, R., Jia, P., and Zhao, Z. (2020). 6mA-Finder: a novel online tool for predicting DNA N6-methyladenine sites in genomes. *Bioinformatics* 36, 3257–3259. doi: 10.1093/bioinformatics/btaa113

Xue, T., Zhang, S., and Qiao, H. (2021). i6mA-VC: A multi-classifier voting method for the computational identification of DNA N6-methyladenine sites. *Interdiscip Sci.* 13, 413–425. doi: 10.1007/s12539-021-00429-4

Yang, H., Luo, Y., Ren, X., Wu, M., He, X., Peng, B., et al. (2021). Risk Prediction of Diabetes: Big data mining with fusion of multifarious physical examination indicators. *Inf. Fusion* 75, 140–149. doi: 10.1016/j.inffus.2021.02.015

Yang, Q., Zhu, W., Tang, X., Wu, Y., Liu, G., Zhao, D., et al. (2024). Improving rice grain shape through upstream ORF editing-mediated translation regulation. *Plant Physiol.* 197, kiae557. doi: 10.1093/plphys/kiae557

Yu, H., and Dai, Z. (2019). SNNRice6mA: A deep learning method for predicting DNA N6-methyladenine sites in rice genome. *Front. Genet.* 10, 1071. doi: 10.3389/fgene.2019.01071

Zhang, H. Q., Arif, M., Thafar, M. A., Albaradei, S., Cai, P., Zhang, Y., et al. (2025). PMPred-AE: a computational model for the detection and interpretation of pathological myopia based on artificial intelligence. *Front. Med. (Lausanne)* 12, 1529335. doi: 10.3389/fmed.2025.1529335

Zhang, G., Huang, H., Liu, D., Cheng, Y., Liu, X., Zhang, W., et al. (2015). N6-methyladenine DNA modification in drosophila. *Cell* 161, 893–906. doi: 10.1016/j.cell.2015.04.018

Zhang, Q., Liang, Z., Cui, X., Ji, C., Li, Y., Zhang, P., et al. (2018). N(6)-methyladenine DNA methylation in japonica and indica rice genomes and its association with gene expression, plant development, and stress responses. *Mol. Plant* 11, 1492–1508. doi: 10.1016/j.molp.2018.11.005

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R. V., and Liu, H. (2023). DNABERT-2: efficient foundation model and benchmark for multi-species genome. *arXiv (USA)*, abs/2306.15006. https://ui.adsabs.harvard.edu/abs/2023arXiv230615006Z/abstract

Zhou, S., Li, X., Liu, Q., Zhao, Y., Jiang, W., Wu, A., et al. (2021). DNA demethylases remodel DNA methylation in rice gametes and zygote and are required for reproduction. *Mol. Plant* 14, 1569–1583. doi: 10.1016/j.molp.2021.06.006

Zhou, C., Wang, C., Liu, H., Zhou, Q., Liu, Q., Guo, Y., et al. (2018). Identification and analysis of adenine N6-methylation sites in the rice genome. *Nat. Plants* 4, 554–563. doi: 10.1038/s41477-018-0214-x

Zhou, Z., Xiao, C., Yin, J., She, J., Duan, H., Liu, C., et al. (2024). PSAC-6mA: 6mA site identifier using self-attention capsule network based on sequence-positioning. *Comput. Biol. Med.* 171, 108129. doi: 10.1016/j.compbiomed.2024.108129

Zhu, S., Beaulaurier, J., Deikus, G., Wu, T. P., Strahl, M., Hao, Z., et al. (2018). Mapping and characterizing N6-methyladenine in eukaryotic genomes using single-molecule real-time sequencing. *Genome Res.* 28, 1067–1078. doi: 10.1101/gr.231068.117

Zhu, H., Hao, H., and Yu, L. (2024). Identification of microbe–disease signed associations via multi-scale variational graph autoencoder based on signed message propagation. *BMC Biol.* 22, 172. doi: 10.1186/s12915-024-01968-0

Zou, K., Wang, S., Wang, Z., Zhang, Z., and Yang, F. (2023). HAR_Locator: a novel protein subcellular location prediction model of immunohistochemistry images based on hybrid attention modules and residual units. *Front. Mol. Biosci.* 10, 1171429. doi: 10.3389/fmolb.2023.1171429

Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. *RNA* 25, 205–218. doi: 10.1261/rna.069112.118

Zou, H. L., Yang, F., and Yin, Z. J. (2022). Integrating multiple sequence features for identifying anticancer peptides. *Comput. Biol. Chem.* 99, 7. doi: 10.1016/j.compbiolchem.2022.107711

Zulfiqar, H., Guo, Z., Ahmad, R. M., Ahmed, Z., Cai, P., Chen, X., et al. (2023). Deep-STP: a deep learning-based approach to predict snake toxin proteins by using word embeddings. *Front. Med. (Lausanne)* 10, 1291352. doi: 10.3389/fmed.2023.1291352