frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 09 June 2025
DOI 10.3389/fpls.2025.1618174

OPEN ACCESS

EDITED BY
Shanwen Sun,
Northeast Forestry University, China

REVIEWED BY
Ran Su,
Tianjin University, China
Xiangzheng Fu,
Hunan University, China

*CORRESPONDENCE
Zhibin Lv
lvzhibin@pku.edu.cn

RECEIVED 25 April 2025
ACCEPTED 12 May 2025
PUBLISHED 09 June 2025

CITATION
Pu Y, Hao X, Zheng Z, Ma H and Lv Z (2025) A BERT-based rice enhancer identification model combined with sequence-representation differential entropy interpretation.
Front. Plant Sci. 16:1618174.
doi: 10.3389/fpls.2025.1618174

A BERT-based rice enhancer identification model combined with sequence-representation differential entropy interpretation

Yajing Pu[1], Xintong Hao[1], Zhaoqi Zheng[1], Huiyan Ma[2] and Zhibin Lv[1]*

[1]College of Biomedical Engineering, Sichuan University, Chengdu, China, [2]College of Life Sciences, Sichuan University, Chengdu, China

Rice is a crucial food crop, and research into its gene expression regulation holds significant importance for molecular breeding and yield improvement. Enhancers, as key elements regulating the spatiotemporal-specific expression of genes, represent a core challenge in functional genomics due to their precise identification requirements. Current deep learning-based methods for rice enhancer identification face limitations primarily in feature extraction efficiency and the generalization capabilities of model architectures. In response, this study introduces a novel model architecture, RiceEN-BERT-SVM, which integrates DNABERT-2 as a feature extraction tool, alongside Support Vector Machine (SVM) for enhancer sequence classification. The mechanism underlying the optimization of model performance is elucidated through differential entropy analysis of feature representations. Experimental results demonstrate the high precision of this approach, achieving an accuracy of 88.05% in 5-fold cross-validation and 87.55% in independent testing. These metrics surpass current state-of-the-art (SOTA) models by margins ranging from 1.47% to 6.87% on the same dataset. Further refinement through fine-tuning enhances RiceEN-BERT-SVM's performance, increasing its accuracy by an additional 6.95%, resulting in a final accuracy of 93.63%. The study employs differential entropy analysis of sequence feature representations to explain the performance enhancements observed with increased fine-tuning iterations. As the number of iterations rises, the differential entropy distributions of positive and negative sample features gradually separate from their initial overlapping state, corresponding with the model's progressive improvement in performance. At six fine-tuning iterations, the separation between positive and negative sample entropy reaches its peak, achieving optimal model performance. Beyond this point, the distributions begin to overlap again, leading to a decline in performance. This novel approach not only offers an efficient tool for rice enhancer identification but also introduces a visually interpretable framework based on differential entropy, providing a new perspective for optimizing biological sequence analysis models.

KEYWORDS

rice enhancer, large language model, positive and negative sample distribution, support vector machine, visual explanation

1 Introduction

An enhancer is a DNA sequence in the genome that can bind to transcription factors and other regulatory proteins to enhance gene transcriptional activity (Sparks et al., 2013; Ding et al., 2023). In the rice genome, enhancers are primarily distributed in the inner regions and near gene loci, playing a crucial role in regulating gene expression (Zhao et al., 2020; Reed et al., 2023; Hamdy et al., 2024; Lin, 2024; Qiao J. et al., 2024; Zhao M. et al., 2024; Zhao Y. et al., 2024; Zhou et al., 2024). The accurate identification of rice enhancers is critical for understanding their biological mechanisms (Cao et al., 2021a; Cao et al., 2021b; Cao et al., 2022). However, traditional identification methods like chromatin immunoprecipitation sequencing (ChIP-seq) (Qiu et al., 2025) and reporter gene experiments, are labor-intensive, inefficient, and lack genome-wide coverage, making it challenging to efficiently identify enhancers throughout the genome.

In terms of computing, especially with the rise of artificial intelligence technology, machine learning-based identification of rice enhancer sequences has garnered increasing attention (Kaur et al., 2019; Machnicka and Wilczynski, 2020; Li et al., 2021; Cheng et al., 2024; Qiao B. et al., 2024; Xie et al., 2024; Yin et al., 2024). Currently, these methods can be categorized into two groups based on the distinct machine learning approaches employed. The first category comprises classic machine learning algorithms reliant on manually designed feature extraction techniques. For instance, Nisha et al. proposed the RFECS (Rajagopal et al., 2013), which integrates multi-omics features (e.g., histone modification and DNA accessibility) using a random forest model, significantly enhancing the accuracy of rice enhancer predictions. Meanwhile, Yinuo et al. introduced iEnhancer-KL (Lyu et al., 2021), combining PSTNPss and Kullback–Leibler (KL) divergence to quantify sequence distribution differences, extract nonlinear features from the rice genome, and ultimately employ SVM for enhancer classification. The second category involves deep neural networks based on automatic feature extraction (Khanal et al., 2020; Gao et al., 2022; Xiao et al., 2025). These approaches focus on improving recognition performance by leveraging various deep neural network architectures. For example, Khanal et al. developed iEnhancer-CNN (Khanal et al., 2020), which integrates word2vec models and convolutional neural networks (CNNs) from natural language processing to identify enhancers directly from raw DNA sequences (Zou et al., 2019). By contrast, Yujia et al. proposed RicENN (Gao et al., 2022), combining CNNs, bidirectional recurrent neural networks (RNNs), and attention mechanisms for the specific recognition of rice enhancers. Although these methods have achieved significant progress in enhancer recognition, they remain constrained by certain limitations. These include challenges related to model interpretability, difficulties in intuitively understanding enhancer regulatory mechanisms, inaccurate feature extraction, and limited generalization across different species.

With the innovation of the Transformer architecture, pre-trained language models have successfully expanded into the field of biomolecular sequence analysis (Liu Y. et al., 2024; Yan K. et al., 2024; Lai et al., 2025). In protein research, models such as ProtTrans (Elnaggar et al., 2022)and ESM series (Rives et al., 2021; Xiao et al., 2024) have been developed to predict protein structure and function. Similarly, in the DNA domain, several improved models based on the BERT architecture (Wei et al., 2020; Le et al., 2021; Ai et al., 2024; Li et al., 2024) exist. For example, the iEnhancer-EL proposed by Liu et al (Liu et al., 2018), employs a multi-scale k-mer labeling strategy to segment DNA sequences and extracts local semantic features of enhancer sequences using a BERT-like framework. However, the k-mer labeling method has limitations: sequence overlap leads to information redundancy, significantly increasing computational complexity, and the selection of k-values requires empirical adjustment, which limits the model's ability to capture long-distance dependencies (Le et al., 2021; Zhou et al., 2023). In contrast, DNABERT-2, a new generation DNA language model, has achieved two major technological breakthroughs. First, it replaces traditional k-mer segmentation with the Byte Pair Encoding (BPE) word segmentation strategy. BPE dynamically merges high-frequency subsequences to generate adaptive tokens. For instance, as demonstrated in Zhihan et al.'s study, BPE encoding reduced tokenized sequence length by a factor of 5 compared to 6-mer tokenization (Zhou et al., 2023), significantly enhancing processing efficiency for long sequences. Second, its Transformer architecture incorporates Attention with Linear Biases (ALiBi) technology (Press et al., 2021), which optimizes position coding and overcomes traditional limitations on input sequence length. DNABERT-2 can flexibly process genomic sequences of any length. Therefore, when dealing with complex DNA sequences of varying lengths, DNABERT-2 demonstrates superior efficiency and accuracy in feature extraction compared to other large language models. It provides a more efficient and powerful tool for genome sequence analysis.

In information theory, information entropy measures the uncertainty and complexity of information. The larger the entropy value, the higher the disorder of the system; this corresponds to richer diversity in possible states and a greater amount of information contained (Shannon, 1948). Differential entropy extends information entropy to continuous random variables and quantifies the uncertainty inherent in their probability distributions (Kozachenko and Leonenko, 1987). If the probability density function f(x) is uniformly distributed, the differential entropy will be larger, indicating higher uncertainty. Conversely, if f(x) is highly concentrated, the entropy may be smaller or even negative. This property allows differential entropy to describe the uncertainty of continuous signals with flexibility, leading to a wide range of applications. In communication engineering, differential entropy can quantify the distribution difference between signals and noise (e.g., calculating the channel capacity limit) (Lapidoth and Moser, 2009), which is essential for optimizing efficient transmission technologies like orthogonal frequency division multiplexing (OFDM) (Li et al., 2007). In physics, differential entropy's mathematical correspondence with thermodynamic entropy (e.g., Boltzmann's entropy formula) provides microscopic probabilistic explanations for the analysis of macroscopic phenomena such as gas diffusion and phase transition (Ellis, 1999; Xing, 2003). In machine learning, mutual information indicators derived from differential entropy overcome the limitations of linear correlation

analysis. For instance, they can capture nonlinear correlations between features and target variables. By measuring these relationships, mutual information facilitates feature screening and model performance improvement (Beirlant et al., 1997; Hanchuan et al., 2005; Leonenko et al., 2008).

To address the limitations in existing methods for rice enhancer sequence prediction regarding feature extraction and model generalization, we proposed a model called RiceEN-BERT-SVM. This model leverages DNABERT-2 as its feature extractor and employs Support Vector Machine (SVM) as the classifier. Experiments demonstrate that RiceEN-BERT-SVM efficiently identifies rice enhancers with an independent test accuracy of 87.55%, surpassing RicENN, the current state-of-the-art model, by 10.82%. By employing DNABERT-2's fine-tuning capability for downstream tasks, we achieved a 93.63% accuracy for RiceEN-BERT-SVM after six fine-tuning iterations. We also utilize differential entropy distributions derived from positive and negative sample features to visually interpret how the performance of our fine-tuning model changes with the number of tuning iterations. Additionally, we propose a method for determining the optimal number of fine-tune iterations by analyzing the polarization of positive and negative sample-averaged differential entropy distances. Our approach not only provides novel tools and ideas for rice enhancer sequence recognition but also offers a fresh perspective on the visual interpretation of model behavior.

2 Materials and methods

2.1 Dataset

To train a rice enhancer prediction model, we obtained genome sequence data of rice (Oryza sativa Japonica Group) from the Ensembl Plants website (Quang et al., 2016; Howe et al., 2020). This database integrates DNA sequence resources from a variety of important crops, providing rich basic data for genomic research. Based on the enhancer active regions verified by Jialei et al. using STARR-seq technology (Sun et al., 2019), 9,642 enhancer sequences were extracted from rice chromosomes as positive sample. Simultaneously, based on DNase I hypersensitivity site (DHS) predictions, we identified 23,398 non-enhancer sequences as negative samples. After redundancy removal with CD-HIT tool (Huang et al., 2010; Li et al., 2012), the final dataset comprised 4,082 enhancers and 9,916 non-enhancers exhibiting a mild class imbalance with a positive-to-negative ratio of 1:2.43. The positive samples were divided into training and test sets in a ratio of 7:3. For the test set, 30% of the positive samples and an equivalent proportion of negative samples were randomly selected to maintain consistency with the original dataset's chromosome distribution. The remaining 70% of both positive and negative samples were allocated for the training set. Ultimately, we constructed a training dataset comprising 9,346 samples and an independent test set consisting of 3,882 samples. During the training process, the training set was further subdivided into five subsets for cross-validation purposes. Of these, 80% was used to train the model, while the remaining 20% served to validate its performance.

2.2 Model architecture

In order to efficiently identify and predict rice enhancers, we propose a new neural network model based on DNABert-2 and explore the fine-tuned performance of the model using differential entropy, as illustrated in Figure 1. This framework comprises three main components: DNABert-2 feature extraction (Zhou et al., 2023), machine learning, and differential entropy computation. When entered into DNABert-2, a DNA sequence generates corresponding feature vectors for both fine-tuned and un-fine-tuned models. Subsequently, these un-fine-tuned feature vectors are inputted into five distinct machine learning algorithms. Each algorithm offers unique advantages in identifying rice enhancers. Results indicate that the SVM algorithm (Hearst et al., 1998; Wang et al., 2024) demonstrates superior performance in classifying rice enhancers. Consequently, we utilize the SVM algorithm to train using the fine-tuned eigenvectors (epoch1-epoch10), thereby obtaining predictions of rice enhancers. Additionally, considering the eigenvector without fine-tuning (epoch0) alongside the 10 epochs of fine-tuning, we estimate differential entropy employing the Kozachenko-Leonenko method (Kozachenko and Leonenko, 1987). This analysis elucidates the impact of fine-tuning through mean and median entropy differences observed in positive and negative samples before and after fine-tuning.

2.2.1 Feature extraction

In DNABert-2, the input layer receives DNA sequences processed through BPE tokenization and embeds them into a high-dimensional space. These embeddings are then passed through 12 Transformer Encoder layers, which serve as the core components responsible for capturing long-distance dependencies and complex patterns within the sequences. Each Transformer Encoder layer incorporates ALiBi (Attention with Linear Biases), a positional encoding mechanism that introduces linearly decaying bias to attention weights based on token distances. This design enables flexible handling of variable-length sequences while maintaining computational efficiency. Additionally, the feedforward layers employ GEGLU (Gated Linear Unit with GELU), which splits the input into two components, applies GELU to one, and multiplies them. This gating mechanism improves nonlinear modeling compared to standard activations (Zhou et al., 2023). In the output layer, a 768-dimensional eigenvector is generated. This un-fine-tuned result from the base architecture is designated as epoch0. If subjected to further fine-tuning and optimization, these eigenvectors can serve as inputs for subsequent tasks. Within this study, the model undergoes a total of 10 training iterations following fine-tuning to yield epochs 1 through 10.

2.2.2 Machine learning

Distinct machine learning algorithms offer various advantages in identifying rice enhancers. We evaluated five common algorithms to preprocess feature vectors derived from Bert: LGBM and XGB, both ensemble methods utilizing gradient boosting, are effective for complex nonlinear relationships, ideal for large datasets, and yield high prediction accuracy (Zou et al., 2023). SVM excels in high-dimensional spaces and small samples by

10.3389/fpls.2025.1618174

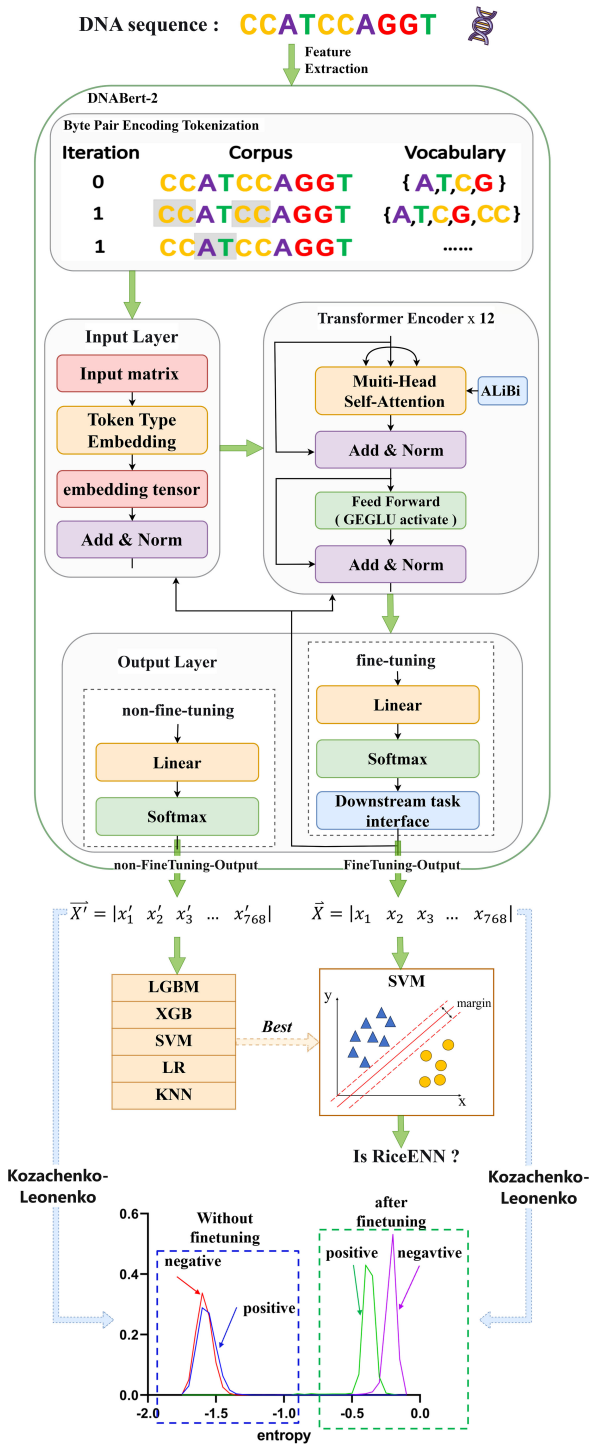

FIGURE 1
Technology roadmap. The DNA sequence was extracted by DNABert-2, and the unfine-tuned and fine-tuned feature vectors were output, respectively. After the feature sequences without fine-tuning were identified by different algorithms, it was found that the SVM algorithm had the best performance, so the features after SVM training were selected to determine whether the enhancer was not. In addition, in order to explain the effect of fine-tuning on the model, the optimal number of fine-tuning was selected, and the differential entropy of positive and negative samples was calculated and visualized.

finding an optimal hyperplane for classification (Zhu et al., 2023). LR is efficient and straightforward, suitable for linearly separable data, providing model interpretability. KNN classifies based on nearest-neighbor similarity, making it apt for tasks with limited data volumes. As shown in Figure 2, SVM outperforms other algorithms across ACC, MCC, Sn, Sp, and Sr metrics for both training and test sets. Consequently, we employed SVM to train the fine-tuned feature vectors. To ascertain the optimal fine-tuning stage for rice enhancer recognition, we input feature vectors from 10 epochs into SVM. This process yielded results for recognizing enhancers from epoch1 to epoch10, facilitating our determination of the best model performance.

2.2.3 Differential entropy calculation

In information theory, differential entropy is used to measure the uncertainty of a continuous random variable, specifically, defined as follows: for a continuous random variable X with probability density function $f(x)$, the differential entropy $h(X)$ (Equation 1) is defined as:

$$h(X) = -\int_{-\infty}^{\infty} f(x)\log f(x)dx \tag{1}$$

This study introduces differential entropy to quantify the ability of DNABert-2 to capture sample features during the learning process, as well as its ability to distinguish between positive and negative samples with varying numbers of fine-tuning times.

However, in practice, the probability density function (PDF) of continuous variables is typically not directly obtainable, necessitating entropy estimation from empirical data. Common approaches include binning-based discretization methods (Alneberg et al., 2014) and kernel density estimation (KDE) (Parzen, 1962). However, binning requires arbitrary discretization of data into intervals, risking information loss or artificial patterns, while KDE suffers from sensitivity to bandwidth selection and high computational costs in high-dimensional spaces. In contrast, the Kozachenko-Leonenko (K-L) entropy estimator circumvents these limitations without requiring explicit PDF estimation (Kozachenko and Leonenko, 1987; Bulinski and Dimitrov, 2019). Specifically, its core principle involves calculating the average distance from each sample point to its k-th nearest neighbor. This approach cleverly bypasses direct density modeling while demonstrating superior efficiency and accuracy in high-dimensional data processing. The formula (Equation 2) is as follows:

$$H_N := d\log\bar{\rho} + \log V_d + \gamma + \log(N-1) \tag{2}$$

Where H_N is the estimate of differential entropy $H(f)$. d is the dimension of the space where the random vector is located. $\bar{\rho}$ is the geometric mean of the nearest neighbor distance in the sample, that is, $\bar{\rho} = (\rho_1 \cdot \rho_2 \cdot \ldots \cdot \rho_N)^{1/N}$, where ρ_i is the distance from the i -th sample point to its nearest neighbor sample point. V_d is the d-dimensional unit sphere volume, that is, $V_d = \frac{\pi^{\frac{D}{2}}}{\Gamma(1+\frac{D}{2})}$. γ is the Euler-Marshalloni constant, which is approximately equal to 0.5772. N is the sample size.

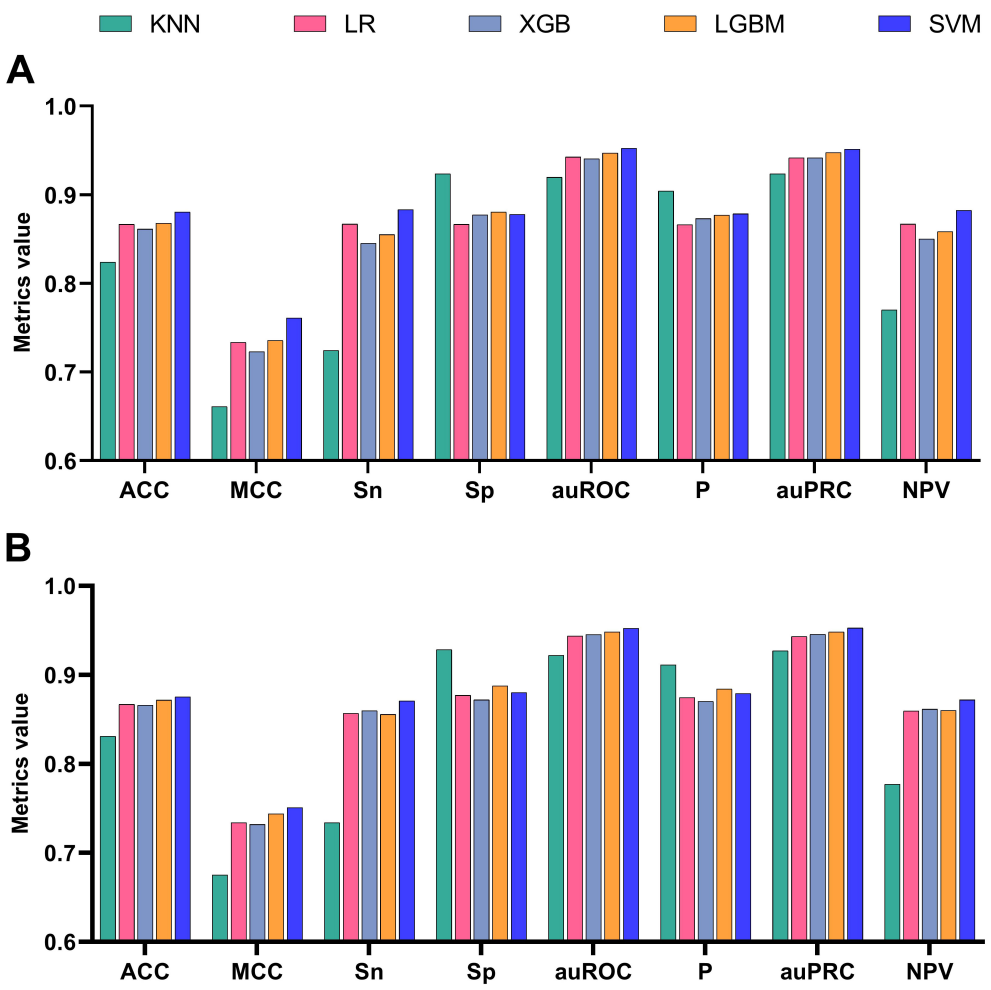

FIGURE 2
Comparison of indicators without fine-tuning for different methods: **(A)** Comparison on the training set, **(B)** Comparison on the test set.

During the fine-tuning process, the differential entropy of the positive and negative samples in each epoch (from 0 to 10) is calculated using Kozachenko–Leonenko, and then the differential entropy of the positive and negative samples changes with the increase in the number of fine-tunings.

2.3 Model evaluation

To comprehensively evaluate the model's performance in rice enhancer recognition, we adopted a 5-fold cross-validation approach combined with independent testing. Based on the training set and test set constructed in the previous section, during the training phase, the training set is divided into 5 subsets. Each time, 4 subsets are used for model training, and the remaining subset is used to validate the model's performance. This process is repeated 5 times (once for each subset as the validation set) to optimize the parameters. Finally, the trained model is tested on the independent test set to obtain the prediction results.

The following metrics were selected to evaluate model performance: accuracy (ACC) (Equation 3), Matthews correlation coefficient (MCC) (Equation 4), recall (Sn) (Equation 5), specificity (Sp) (Equation 6), negative predictive value (NPV) (Equation 7), precision (P) (Equation 8), auROC, and auPRC (Grau et al., 2015; Liu M. et al., 2024; Zhu et al., 2024; Huang et al., 2025; Zhang et al., 2025). These metrics measure the classification performance of the model from different perspectives, and they are defined below:

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

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

$$Sn = \frac{TP}{TP + FN} \tag{5}$$

$$Sp = \frac{TN}{TN + FP} \tag{6}$$

$$NPV = \frac{TN}{TN + FN} \quad (7)$$

$$P = \frac{TP}{TP + FP} \quad (8)$$

Where TP represents true positive (TP), TN represents true negative (TN), FP represents false positive (FP), and FN represents false negative (FN). auROC stands for area under the ROC curve, which plots recall (Sn) and false positive rate (FPR) at different thresholds. Values closer to 1 indicate better performance. Similarly, auPRC stands for area under the precision-recall curve, which plots precision (P) and recall (Recall) at different thresholds; higher values closer to 1 indicate better performance.

3 Results and discussions

3.1 Analyzing ML models with pretrained LM feature extraction

The DNABERT-2 large language model employs a dynamic compression algorithm in place of the fixed k-mer window and implements the ALiBi attention mechanism to facilitate full-sequence modeling. Following extensive training on big data, it effectively captures potential feature information within sequences. Leveraging this capability, DNABERT-2 was applied to feature extraction in rice enhancer recognition tasks to evaluate its practical effectiveness in identifying rice enhancers. Simultaneously, the extracted features were entered into five distinct machine learning algorithms for classification purposes; their performances were then compared to identify the most suitable algorithm for future fine-tuning.

The experimental findings are illustrated in Figure 2. Figure A presents a performance comparison of the five machine learning algorithms during cross-validation, while Figure B evaluates them under independent testing conditions. It is evident that SVM emerges as the top performer across six out of eight metrics, excluding Sp and P, demonstrating consistent superiority irrespective of whether it was trained on or tested against datasets. In cross-validation, the SVM algorithm achieved a peak performance of 0.883, whereas in independent testing, its score reached 0.879. The outcomes of all eight metrics were averaged across both validation methods, reinforcing SVM's superior performance. Compared to the other four algorithms, SVM exhibited improvements ranging from 1.41% to 6.27% during cross-validation and 0.49% to 5% during independent testing. Consequently, the SVM algorithm was identified as the most effective for the rice enhancer recognition task and selected as the foundational model for subsequent fine-tuning exercises.

3.2 Model fine-tuning effects

DNABert-2 can be fine-tuned for the downstream task of identifying rice enhancers. The model was fine-tuned for 10 iterations, with each fine-tuning run denoted as Epoch_i (i = 1, 2, …, 10), while the unmodified baseline was labeled Epoch_0. To mitigate overfitting risks, we prioritized evaluation on an independent test set. The performance of all 11 epochs (Epoch_0 to Epoch_10) was systematically compared, and the results are shown in Figure 3. Among the 8 evaluation metrics, AUC (Area Under the ROC Curve) was selected as the primary metric due to its threshold independence, which comprehensively aggregates model performance across all classification thresholds without requiring arbitrary cutoff selection. Additionally, AUC directly quantifies the model's ranking ability, aligning with the practical need for enhancer identification in genomic studies.

Figure 3A demonstrates progressive performance improvement through fine-tuning at the macro scale (0–1 auROC range), while the magnified view in Figure 3B (0.950-0.980 auROC range) reveals performance oscillations. Our analysis suggests no clear correlation exists between model performance and the number of fine-tuning iterations, as the performance improvement does not scale linearly

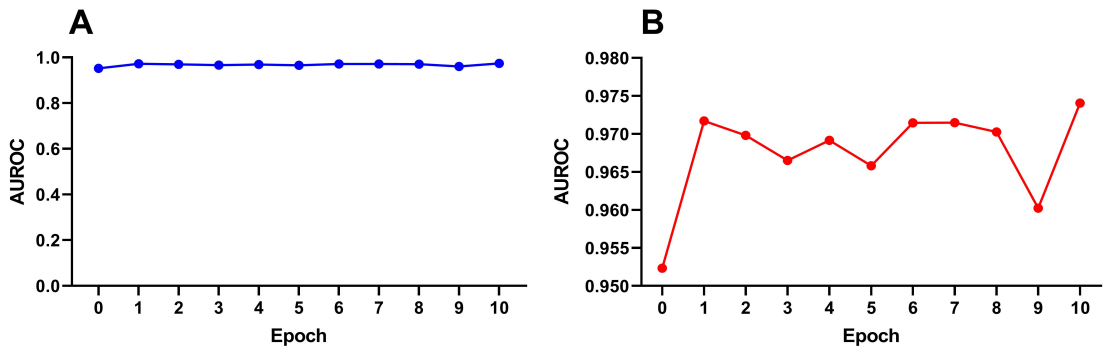

FIGURE 3
Variation trend of auROC in the testset with the number of fine-tuning times under SVM. **(A)** Full-scale view (y-axis: 0–1) demonstrating performance improvement through fine-tuning. **(B)** Magnified view (y-axis: 0.950–0.980) highlighting nuanced auROC fluctuations.

10.3389/fpls.2025.1618174

with additional fine-tuning epochs. During the first fine-tuning step (Epoch_1), there was a significant improvement in performance compared to Epoch_0. After that, the AUC fluctuated within a narrow range of 0.965–0.975 until Epoch 9, after which the AUC dropped significantly. It is hypothesized that excessive fine-tuning at this stage may lead to model overfitting, thereby reducing its generalization ability on the test set. At this point, we cannot definitively determine the optimal number of fine-tuning epochs.

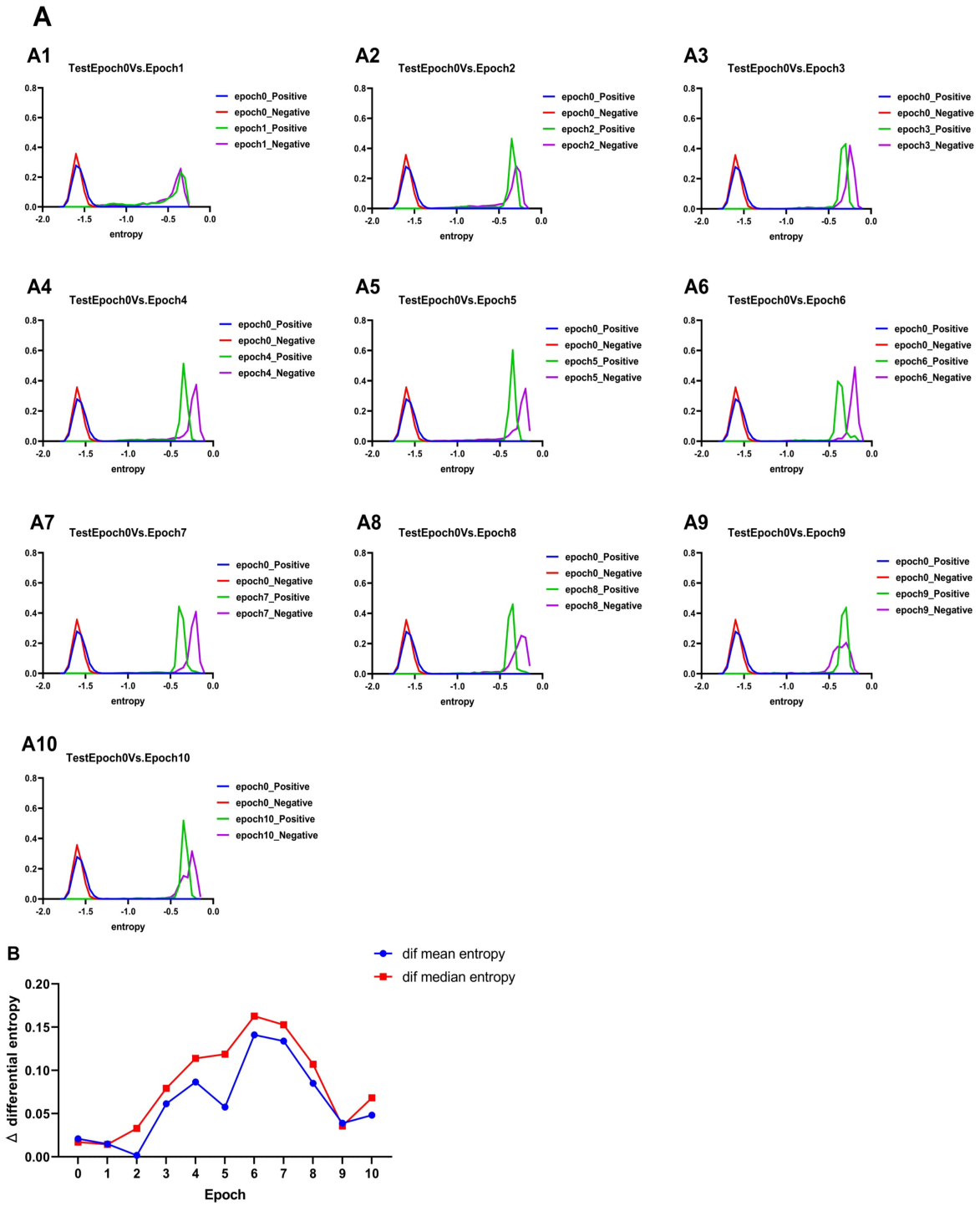

FIGURE 4
Results of differential entropy. **(A)** Change in entropy of the test set. **(B)** Change of the mean entropy difference and median entropy difference of positive and negative samples in the test set with the increase of epoches (0–10 eopoches). Where Δmean differential entropy =Positive sample mean differential entropy - Negative sample mean differential entropy |, Δmedian differential entropy =Positive sample median differential entropy - Negative sample median differential entropy |.

3.3 Differential entropy explanation

We calculate differential entropy to more intuitively explain the effect of fine-tuning on the model. Differential entropy measures the uncertainty of data distributions and can be used to evaluate the model's ability to distinguish between positive and negative samples. Specifically, we used the Kozachenko–Leonenko estimator to calculate the differential entropy of the positive and negative samples for Epoch_i(i=0,1,2… 10), visualizing them to observe the change patterns. As shown in Figure 4A, with an increase in the number of fine-tuning epochs, the differential entropy of positive and negative samples exhibits a trend of "coincidence–separation–coincidence." At the initial stage, the differential entropy of positive and negative samples coincided, indicating that the model had not yet fully distinguished between them. During the fine-tuning process, the differential entropy of positive and negative samples gradually separated, suggesting that the model's ability to differentiate improved. Notably, during Epoch_6–Epoch_8, the positive and negative samples were most distinct in terms of differential entropy, which also corresponds to the "plateau" phase observed for Epoch_6–Epoch_8 in Figure 3. In the later stages, the differential entropy of positive and negative samples began to coincide again, likely because the model started overfitting and lost its generalization ability. At this point, it can be preliminarily determined that there is an optimal number of fine-tuning epochs between Epoch_6–Epoch_8.

To determine the optimal number of fine-tuning epochs, we calculated the mean and median differential entropy of positive and negative samples in each cycle. These values were used to compute the mean entropy difference and median entropy difference for Epoch_i, which are presented in Figure 4B. The results indicate that the average and median entropy differences between positive and negative samples achieve their maximum at the 6th training cycle (Epoch_6). This suggests that the model exhibits its strongest ability to distinguish between positive and negative samples during this cycle. Based on these findings, Epoch_6 was identified as the optimal model for extracting rice enhancer features.

TABLE 1 Comparison of our methods with other methods on the independent testset.

Method	ACC	SP	SN/REC	PRE	NPV	AUPRC	AUROC
iEnhancer-EL	0.567	0.463	0.671	0.555	0.584	0.695	0.567
iEnhancer-CNN	0.571	0.459	0.682	0.558	0.591	0.700	0.571
RicENN	0.790	0.793	0.788	0.792	0.789	0.879	0.877
RiceEN-BERT-SVM(non-FT)	0.875	0.880	0.871	0.879	0.872	0.952	0.953
RiceEN-BERT-SVM(FT)	**0.936**	**0.932**	**0.941**	**0.932**	**0.940**	**0.953**	**0.971**

Bold values are the models that achieve the best performance.

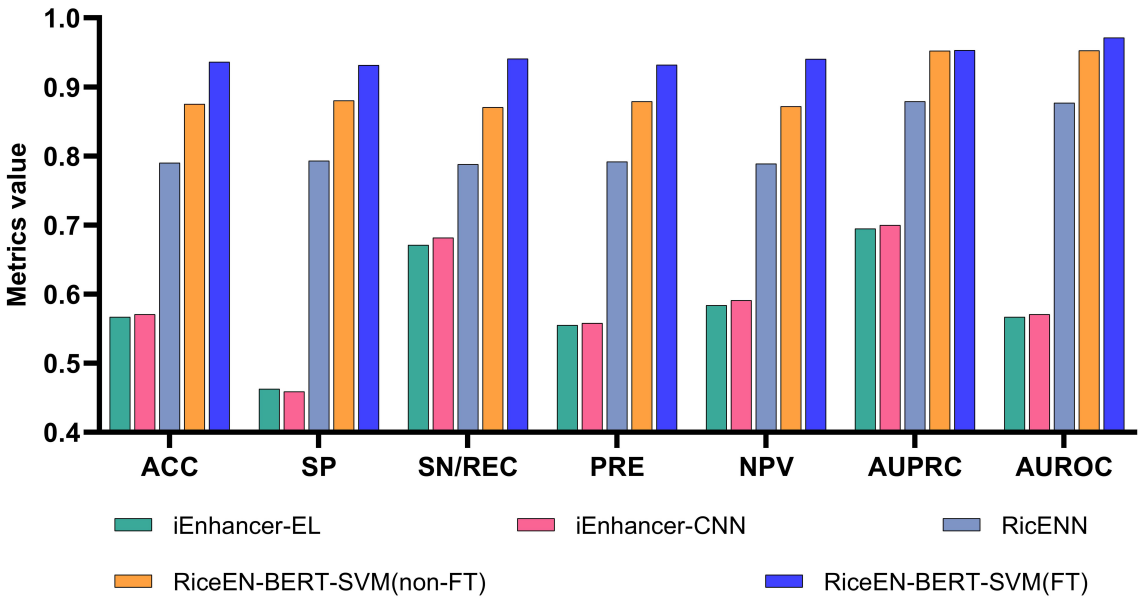

FIGURE 5
This study is compared with other methods. Our model is RiceEN-BERT-SVM, non-FT means that the model is not fine-tuned, and FT means the fine-tuned model.

3.4 Comparisons with the existing methods

Based on the best model selected above, as well as the previously unfine-tuned RiceEN-BERT-SVM, we compared our models with other existing rice enhancer recognition methods on an independent test set, including RicENN (Gao et al., 2022), iEnhancer-CNN (Khanal et al., 2020), and iEnhancer-EL (Liu et al., 2018). Table 1 and Figure 5 present the comparative performance metrics between our framework and state-of-the-art methods. The results demonstrate that both model configurations of RiceEN-BERT-SVM —the fine-tuned (6 iterations) and baseline (unfine-tuned) versions— consistently surpass all existing approaches across evaluation metrics. They achieved first and second places in all 7 evaluation criteria, with significant improvements in performance. Notably, the fine-tuned RiceEN-BERT-SVM demonstrated exceptional average performance during independent testing, with specific metric values as follows: ACC, SP, SN/REC PRE, NPV, AUPRC, and AUROC were 0.936, 0.932, 0.941, 0.932, 0.940, 0.953, and 0.971, respectively. Compared to RicENN, which ranked third, the performance metrics for our fine-tuned model improved by 8.42% to 19.42%. These results clearly demonstrate that the fine-tuned Bert-2 framework is effective in recognizing rice enhancers.

4 Conclusion

We used DNABERT-2, a pre-trained large language model for biological sequences, to extract features from rice enhancer sequences. Combining these features with a support vector machine (SVM) classifier, we constructed a novel model, RiceEN-BERT-SVM, designed to identify rice enhancers. The model demonstrated exceptional performance, achieving cross-validation and independent test results that significantly outperformed existing state-of-the-art methods. Specifically, on the independent test set, our model achieved an accuracy (ACC) of 93.63% and an area under the receiver operating characteristic curve (AUROC) of 97.15%. These metrics represent improvements of 18.52% and 10.77%, respectively, compared to RicENN. To further understand model performance, we developed a methodology to visualize the discriminative ability of the model by leveraging the differential entropy representation of large-language-embedding features derived from DNA sequences. During fine-tuning experiments, we observed that with an increasing number of fine-tuning iterations, the differential entropy between positive and negative samples initially separated and then converged. This trend indicated that the model's discriminative capacity first increased and later weakened as fine-tuning progressed. At a specific fine-tuning threshold (6 iterations), the difference in differential entropy between positive and negative samples was maximized, coinciding with peak model performance. Our findings demonstrate two key insights: First, pre-trained large language models like DNABERT-2 can significantly enhance the recognition of rice enhancer sequences. Second, the changes in fine-tuning performance are closely tied to shifts in the representation of positive and negative sample distributions, as captured by differential entropy. The approach of visualizing differential entropy in feature representations is broadly applicable and can serve as a valuable tool in future studies involving machine learning for DNA, RNA, or protein sequence recognition. However, our study has limitations. For instance, the differential entropy analysis does not account for spatial patterns in DNA sequences, potentially overlooking biologically meaningful structural dependencies. Additionally, the computational efficiency of the current framework can be further optimized. Future studies could develop entropy metrics integrating spatial sequence context and design lightweight architectures to enable broader genomic applications, thereby advancing efficient computational tools for crop molecular design.

Data availability statement

The raw sequence data used in the study were obtained from the following URL: https://plants.ensembl.org/index.html.

Author contributions

YP: Formal Analysis, Validation, Visualization, Writing – original draft. XH: Writing – original draft. ZZ: Writing – original draft. HM: Writing – original draft. ZL: Conceptualization, Funding acquisition, 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 was supported by the National Natural Science Foundation of China (Nos. 62371318 and 62001090), and the 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

Ai, C., Yang, H., Liu, X., Dong, R., Ding, Y., and Guo, F. (2024). MTMol-GPT: *De novo* multi-target molecular generation with transformer-based generative adversarial imitation learning. *PloS Comput. Biol.* 20, e1012229. doi: 10.1371/journal.pcbi.1012229

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., et al. (2014). Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11, 1144–1146. doi: 10.1038/nmeth.3103

Beirlant, J., Dudewicz, E., Gyor, L., and Meulen, E. C. (1997). Nonparametric entropy estimation: an overview. *Int. J. Math. Stat. Sci.* 6.

Bulinski, A., and Dimitrov, D. (2019). Statistical estimation of the shannon entropy. *Acta Mathematica Sinica-English Ser.* 35, 17–46. doi: 10.1007/s10114-018-7440-z

Cao, C., Ding, B., Li, Q., Kwok, D., Wu, J., and Long, Q. (2021a). Power analysis of transcriptome-wide association study: Implications for practical protocol choice. *PloS Genet.* 17, e1009405. doi: 10.1371/journal.pgen.1009405

Cao, C., He, J., Mak, L., Perera, D., Kwok, D., Wang, J., et al. (2021b). Reconstruction of microbial haplotypes by integration of statistical and physical linkage in scaffolding. *Mol. Biol. Evol.* 38, 2660–2672. doi: 10.1093/molbev/msab037

Cao, C., Wang, J., Kwok, D., Cui, F., Zhang, Z., Zhao, D., et al. (2022). webTWAS: a resource for disease candidate susceptibility genes identified by transcriptome-wide association study. *Nucleic Acids Res.* 50, D1123–d1130. doi: 10.1093/nar/gkab957

Cheng, H., Ding, S., and Jia, C. (2024). Prediction of super-enhancers based on mean-shift undersampling. *Curr. Bioinf.* 19, 651–662. doi: 10.2174/0115748936268302231110111456

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

Ellis, R. S. (1999). The theory of large deviations: from Boltzmann's 1877 calculation to equilibrium macrostates in 2D turbulence. *Physica D-Nonlinear Phenomena* 133, 106–136. doi: 10.1016/s0167-2789(99)00101-3

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., et al. (2022). ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 7112–7127. doi: 10.1109/tpami.2021.3095381

Gao, Y. J., Chen, Y. Q., Feng, H. S., Zhang, Y. H., and Yue, Z. Y. (2022). RicENN: prediction of rice enhancers with neural network based on DNA sequences. *Interdiscip. Sciences-Computational Life Sci.* 14, 555–565. doi: 10.1007/s12539-022-00503-5

Grau, J., Grosse, I., and Keilwagen, J. (2015). PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* 31, 2595–2597. doi: 10.1093/bioinformatics/btv153

Hamdy, R., Omar, Y., and Maghraby, F. (2024). DeepEpi: deep learning model for predicting gene expression regulation based on epigenetic histone modifications. *Curr. Bioinf.* 19, 624–640. doi: 10.2174/1574893618666230818121046

Hanchuan, P., Fuhui, L., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1226–1238. doi: 10.1109/TPAMI.2005.159

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Syst. Their Appl.* 13, 18–28. doi: 10.1109/5254.708428

Howe, K. L., Contreras-Moreira, B., De Silva, N., Maslen, G., Akanni, W., Allen, J., et al. (2020). Ensembl Genomes 2020-enabling non-vertebrate genomic research. *Nucleic Acids Res.* 48, D689–D695. doi: 10.1093/nar/gkz890

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26, 680–682. doi: 10.1093/bioinformatics/btq003

Huang, Z., Xiao, Z., Ao, C., Guan, L., and Yu, L. (2025). Computational approaches for predicting drug-disease associations: a comprehensive review. *Front. Comput. Sci.* 19, 1–15. doi: 10.1007/s11704-024-40072-y

Kaur, A., Chauhan, A. P. S., Aggarwal, A. K., and Ieee. (2019). "Machine learning based comparative analysis of methods for enhancer prediction in genomic data," in *2nd International Conference on Intelligent Communication and Computational Techniques (ICCT)*, (India Jaipur: Manipal Univ Jaipur, Jaipur, INDIA, Sep 28-29 2019. 142–145.

Khanal, J., Tayara, H., and Chong, K. T. (2020). Identifying enhancers and their strength by the integration of word embedding and convolution neural network. *IEEE Access* 8, 58369–58376. doi: 10.1109/access.2020.2982666

Kozachenko, L. F., and Leonenko, N. N. (1987). Sample estimate of the entropy of a random vector. *Probl. Inf. Transm. (USA)* 23, 95–101.

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, 155. doi: 10.1145/3714455

Lapidoth, A., and Moser, S. M. (2009). On the capacity of the discrete-time poisson channel. *IEEE Trans. Inf. Theory* 55, 303–322. doi: 10.1109/tit.2008.2008121

Le, N. Q. K., Ho, Q.-T., Nguyen, T.-T.-D., and Ou, Y.-Y. (2021). A transformer architecture based on BERT and 2D convolutional neural network to identify DNA enhancers from sequence information. *Briefings Bioinf.* 22. doi: 10.1093/bib/bbab005

Leonenko, N., Pronzato, L., and Savani, V. (2008). *A class of Rényi information estimators for multidimensional densities*. Available: https://ui.adsabs.harvard.edu/abs/2008arXiv0810.5302L (Accessed October 1, 2008).

Li, W., Fu, L., Niu, B., Wu, S., and Wooley, J. (2012). Ultrafast clustering algorithms for metagenomic sequence analysis. *Briefings Bioinf.* 13, 656–668. doi: 10.1093/bib/bbs035

Li, X., Mardling, R., Armstrong, J., and Ieee. (2007). "Channel capacity of IM/DD optical communication systems and of ACO-OFDM," in *IEEE International Conference on Communications (ICC 2007)*, (Glasgow, Scotland), 2128.

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, 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

Lin, H. (2024). Artificial intelligence with great potential in medical informatics: A brief review. *Medinformatics* 1, 2–9. doi: 10.47852/bonviewMEDIN42022204

Liu, M., Li, C., Chen, R., Cao, D., and Zeng, X. (2024). Geometric deep learning for drug discovery. *Expert Syst. Appl.* 240, 122498. doi: 10.1016/j.eswa.2023.122498

Liu, B., Li, K., Huang, D.-S., and Chou, K.-C. (2018). iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach. *Bioinformatics* 34, 3835–3842. doi: 10.1093/bioinformatics/bty458

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

Lyu, Y., Zhang, Z., Li, J., He, W., Ding, Y., and Guo, F. (2021). iEnhancer-KL: A novel two-layer predictor for identifying enhancers by position specific of nucleotide composition. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 18, 2809–2815. doi: 10.1109/TCBB.2021.3053608

Machnicka, M. A., and Wilczynski, B. (2020). "Chapter 9 - Machine learning and deep learning for the advancement of epigenomics," in *Epigenetics of the Immune System*. Eds. D. Kabelitz and J. Bhat (Academic Press), 217–237.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Stat* 33, 1065–1076, 1012. doi: 10.1214/aoms/1177704472

Press, O., Smith, N. A., and Lewis, M. (2021). Train short, test long: attention with linear biases enables input length extrapolation. Available online at: https://ui.adsabs.harvard.edu/abs/2021arXiv210812409P (Accessed August 1, 2021).

Qiao, J., Jin, J., Yu, H., and Wei, L. (2024). Towards retraining-free RNA modification prediction with incremental learning. *Inf. Sci.* 660, 120105. doi: 10.1016/j.ins.2024.120105

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. doi: 10.1093/bioinformatics/btae581

Qiu, Y., Liu, L., Yan, J., Xiang, X., Wang, S., Luo, Y., et al. (2025). Precise engineering of gene expression by editing plasticity. *Genome Biol.* 26, 51. doi: 10.1186/s13059-025-03516-7

Quang, O., Phuc, N., Nguyen Phuong, T., and Ly, L. (2016). Bioinformatics approach in plant genomic research. *Curr. Genomics* 17, 368–378. doi: 10.2174/1389202917666160331202956

Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., et al. (2013). RFECS: A random-forest based algorithm for enhancer identification from chromatin state. *PloS Comput. Biol.* 9, e1002968. doi: 10.1371/journal.pcbi.1002968

Reed, E., Ferrari, E., and Soloviev, M. (2023). Quality control of gene expression data allows accurate quantification of differentially expressed biological pathways. *Curr. Bioinf.* 18, 409–427. doi: 10.2174/1574893618666230221141815

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. doi: 10.1073/pnas.2016239118

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Sparks, E., Wachsman, G., and Benfey, P. N. (2013). Spatiotemporal signalling in plant development. *Nat. Rev. Genet.* 14, 631–644. doi: 10.1038/nrg3541

Sun, J., He, N., Niu, L., Huang, N., Shen, W., Zhang, Y., et al. (2019). Global quantitative mapping of enhancers in rice by STARR-seq. *Genomics Proteomics Bioinf.* 17, 140–153. doi: 10.1016/j.gpb.2018.11.003

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. (2020). Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework. *Briefings Bioinf.* 22. doi: 10.1093/bib/bbaa275

Xiao, Z., Li, Y., Ding, Y., and Yu, L. (2025). EPIPDLF: a pre-trained deep learning framework for predicting enhancer-promoter interactions. *Bioinformatics* 41, btae716. doi: 10.1093/bioinformatics/btae716

Xiao, C., Zhou, Z., She, J., Yin, J., Cui, F., and Zhang, Z. (2024). PEL-PVP: Application of plant vacuolar protein discriminator based on PEFT ESM-2 and bilayer LSTM in an unbalanced dataset. *Int. J. Biol. Macromol.* 277, 134317. doi: 10.1016/j.ijbiomac.2024.134317

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. doi: 10.1093/bib/bbae495

Xing, X. S. (2003). On the formula for entropy production rate. *Acta Physica Sin.* 52, 2969–2977. doi: 10.7498/aps.52.2970

Yan, K., Lv, H., Shao, J., Chen, S., and Liu, B. (2024). TPpred-SC: multi-functional therapeutic peptideprediction based on multi-label supervisedcontrastive learning. *Sci. China Inf. Sci.* 67, 212105. doi: 10.1007/s11432-024-4147-8

Yin, C., Wang, R., Qiao, J., Shi, H., Duan, H., Jiang, X., et al. (2024). NanoCon: contrastive learning-based deep hybrid network for nanopore methylation detection. *Bioinformatics* 40, btae046. doi: 10.1093/bioinformatics/btae046

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.* 12, 1529335. doi: 10.3389/fmed.2025.1529335

Zhao, Y., Gui, L., Hou, C., Zhang, D., and Sun, S. (2024). GwasWA: A GWAS one-stop analysis platform from WGS data to variant effect assessment. *Comput. Biol. Med.* 169, 107820. doi: 10.1016/j.compbiomed.2023.107820

Zhao, M., Li, J., Liu, X., Ma, K., Tang, J., and Guo, F. (2024). A gene regulatory network-aware graph learning method for cell identity annotation in single-cell RNA-seq data. *Genome Res.* 34, 1036–1051. doi: 10.1101/gr.278439.123

Zhao, L., Xie, L., Zhang, Q., Ouyang, W., Deng, L., Guan, P., et al. (2020). Integrative analysis of reference epigenomes in 20 rice varieties. *Nat. Commun.* 11, 2658. doi: 10.1038/s41467-020-16457-5

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., and Liu, H. (2023). *DNABERT-2: Efficient foundation model and benchmark for multi-species genome*. Available online at: https://ui.adsabs.harvard.edu/abs/2023arXiv230615006Z (Accessed June 1, 2023).

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, 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

Zhu, W., Yuan, S. S., Li, J., Huang, C. B., Lin, H., and Liao, B. (2023). A first computational frame for recognizing heparin-binding protein. *Diagn. (Basel)* 13. doi: 10.3390/diagnostics13142465

Zou, X., Ren, L., Cai, P., Zhang, Y., Ding, H., Deng, K., et al. (2023). Accurately identifying hemagglutinin using sequence information and machine learning methods. *Front. Med. (Lausanne)* 10, 1281880. doi: 10.3389/fmed.2023.1281880

Zou, Q., Xing, P., Wei, L., and Liu, B. (2018). Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. *RNA* 25, 205–218. doi: 10.1261/rna.069112.118