Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MLC-KCR: An imbalance-aware deep learning framework for multi-species plant lysine crotonylation site prediction

Huachun Zhang
Jingdezhen Ceramic University
Jianhua Jia
jjh163yx@163.com
Jingdezhen Ceramic University
Hui Huang
Jingdezhen Ceramic University

Research Article

Keywords: Lysine crotonylation, Deep learning, Multi-species joint training, Class imbalance

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11722915/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MLC-KCR: An imbalance-aware deep learning framework for multi-species plant lysine crotonylation site prediction

Huachun Zhang[1], Jianhua Jia[1]*, Hui Huang[1]

[1]School of Information Engineering, Jingdezhen Ceramic University, 333403 Jingdezhen, Jiangxi, China

*Correspondence: jjh163yx@163.com (J. Jia)

MLC-KCR: An imbalance-aware deep learning framework for multi-species plant lysine crotonylation site prediction

Abstract

Background: Lysine crotonylation (Kcr) is a crucial post-translational modification in plants. However, existing computational predictors are constrained by limited feature representations, inherent class imbalance, and insufficient generalization, which hinder predictive accuracy. To address these challenges, this study proposes MLC-KCR, an accurate and robust deep learning framework designed for plant non-histone Kcr site prediction.

Results: The proposed MLC-KCR integrates local and global feature encodings to capture comprehensive sequence information. Its cascaded architecture employs a multi-scale convolutional network to extract local motifs, bidirectional long short-term memory to model long-range dependencies, and a convolutional block attention module to recalibrate feature weights. To mitigate class imbalance, FinetuneBinaryLoss is combined with an ensemble strategy to enhance minority-class learning and reduce prediction variance. Evaluations on the independent test set demonstrate that MLC-KCR achieved average scores of 81.9% for Sn, 84.8% for Sp, 84.2% for Acc, 68.6% for F1-score, 59.8% for MCC, and 91.0% for AUC, outperforming five existing methods. Furthermore, multi-species joint training outperformed species-specific models, demonstrating shared modification patterns across different plant species. A user-friendly web server was also developed to facilitate predictions without requiring users to specify plant species.

Conclusions: MLC-KCR provides robust predictive performance and serves as a practical tool for plant Kcr site identification. Its implementation offers valuable insights into modification mechanisms and advances plant biology research.

Keywords: Lysine crotonylation; Deep learning; Multi-species joint training; Class imbalance

1. Introduction

Protein post-translational modifications (PTMs) occur dynamically during the late stages of protein biosynthesis. By covalently adding chemical groups to specific amino acid residues, PTMs modify the physicochemical properties, structure, and function of proteins [1]. These modifications are essential for regulating various biological processes, including signal transduction [2], metabolism [3], gene activation [4], and cell differentiation [5]. Recently, advances in high-resolution mass spectrometry have enabled the discovery of novel PTMs [6]. Among these, lysine

crotonylation (Kcr), an evolutionarily conserved acylation modification, has gained considerable attention since its discovery in mammalian histones [7]. Initial research primarily focused on Kcr within histone proteins. Histone Kcr is strongly associated with active promoters and enhancers, functioning as a key epigenetic mark for transcription [8]. However, subsequent proteomic studies revealed that Kcr is not limited to histones. It is widely distributed across various non-histone proteins in both animals and plants. In plants, global identification of non-histone Kcr has been achieved in several major species, including rice [9], wheat [10], tobacco [11], papaya [12], and peanut [13].

Currently, identifying Kcr sites relies heavily on conventional biochemical techniques, such as high-performance liquid chromatography (HPLC) and liquid chromatography-tandem mass spectrometry (LC-MS/MS) [14]. While accurate, these experimental methods are labor-intensive, time-consuming, and expensive [15]. These limitations restrict their application in large-scale or proteome-wide analyses. Consequently, there is an urgent need for efficient and cost-effective computational approaches to rapidly screen putative Kcr sites.

Over the past decade, numerous computational predictors for Kcr sites have been developed. Early models focused exclusively on histone proteins, utilizing traditional machine learning algorithms combined with hand-crafted features. Notable examples include CKSAAP_CrotSite [16], iKcr-PseEns [17], LightGBM-CroSite [18], and iCrotoK-PseAAC [19]. Following the discovery of Kcr in non-histones, researchers began training models on mixed datasets of histones and non-histones. Deep learning architectures substantially improved predictions on these mixed datasets. Deep-Kcr [20] adopted CNNs to extract local motifs. Adapt-Kcr [21] introduced an adaptive embedding framework, while DeepCap-Kcr utilized capsule networks to model hierarchical spatial relationships. Recently, pre-trained language models, such as BERT in BERT-Kcr [22] and ProtT5 in LMCrot[23], were applied to capture deep semantic information. Furthermore, comprehensive frameworks like COMPASS-PTM [24] explored overarching PTM crosstalk. However, mixing histones and non-histones during training often introduces bias due to their distinct sequence characteristics. To address this, recent predictors have been specifically tailored for non-histone proteins.

NhKcr [15] and iKcr_CNN [25] utilized CNNs for human non-histone Kcr prediction. Advanced architectures, such as DenseNet with channel attention in iKcr-DRC [26] and ESM-2 with deep supervision in EDS-Kcr [27], further improved feature extraction. DeepKcrot [28] and CapsNh-Kcr [29] expanded their scope to multi-species prediction. Despite this progress, a notable gap remains: almost all of these non-histone predictors were designed for human or mammalian data. Due to evolutionary divergence, models trained on mammalian sequences generally show poor generalization when applied to plant proteins.

To date, predicting Kcr sites specifically on plant non-histone proteins remains largely unexplored. Recently, PlantNh-Kcr [14] emerged as a pioneering model tailored specifically for identifying Kcr sites in plants, utilizing a framework combining CNN, BiLSTM, and an attention mechanism. While it represents a substantial step forward, the field of plant non-histone Kcr prediction still faces notable bottlenecks that constrain model performance and generalization. First, regarding feature representation, models such as PlantNh-Kcr rely predominantly on single encoding types like binary encoding or basic physicochemical properties. These shallow representations suffer from low dimensionality and fail to fully exploit the deep semantics and long-range dependencies of amino acid residues within complex folding environments. Second, in real-world biological sequences, unmodified lysines vastly outnumber modified ones, creating a notable class imbalance. Currently, mainstream solutions to this pervasive issue focus on the data and algorithm levels, yet both exhibit notable limitations. At the data level, existing predictors typically rely on simple resampling strategies such as random under-sampling. This static preprocessing, however, inevitably leads to the permanent loss of key feature information hidden within the discarded negative samples. At the algorithm level, when employing conventional static loss functions, the optimization process is easily dominated by the gradients of the overwhelming majority class, resulting in exceptionally low sensitivity toward genuine modification sites. Compounding these issues, a single deep neural network trained on highly imbalanced distributions often exhibits high predictive variance and instability, making it prone to getting trapped in local optima.

To address the aforementioned challenges, we propose MLC-KCR, an explainable and robust deep learning framework for plant non-histone lysine crotonylation site prediction. The major contributions of this study are:

1. Deep Integration of Multi-Source Features: To address the semantic integration deficiency of existing models, we construct a comprehensive feature representation by fusing three complementary encodings. We leverage the pre-trained ESM-2 language model to extract deep global semantics, apply One-hot encoding to provide precise positional information, and incorporate the BLOSUM62 matrix to introduce species-evolutionary conservation. This complementarity of "global semantics and local features" lays a solid data foundation for extracting highly discriminative features in subsequent deep networks.

2. Adaptive Loss and Ensemble Generalization Mechanism: To tackle the class imbalance inherent in non-histone datasets, we introduce a synergistic optimization mechanism. Specifically, we employ the FinetuneBinaryLoss [24] a dynamically adaptive focal loss function, to optimize the training process. By computing the real-time ratio of positive to negative samples within each mini-batch, it automatically adjusts the weight α, substantially reducing the computational burden of manual parameter searching and effectively smoothing out local gradient jitters caused by random sampling. Complementing this, we implement a cross-validation-based homogeneous ensemble strategy for the final prediction. By aggregating the predictive probabilities from multiple trained models, this mechanism reduces prediction variance and enhances the robustness of the final decision boundary against data stochasticity.

3. Adaptive Multi-Scale Modeling: We construct a composite deep learning architecture combining Multi-scale CNN, BiLSTM, and CBAM. This synergistic design effectively models both local structural motifs and long-range sequential dependencies, while dynamically recalibrating feature importance along spatial and channel dimensions.

4. User-Friendly Web Server: We develop an online platform that enables convenient submission of protein sequences and retrieval of predictions without computational barriers.

The overall architecture of MLC-KCR is illustrated in Figure 1. This study provides a convenient and robust tool for proteomics analysis and lays a solid foundation for further research on the mechanism of lysine crotonylation.

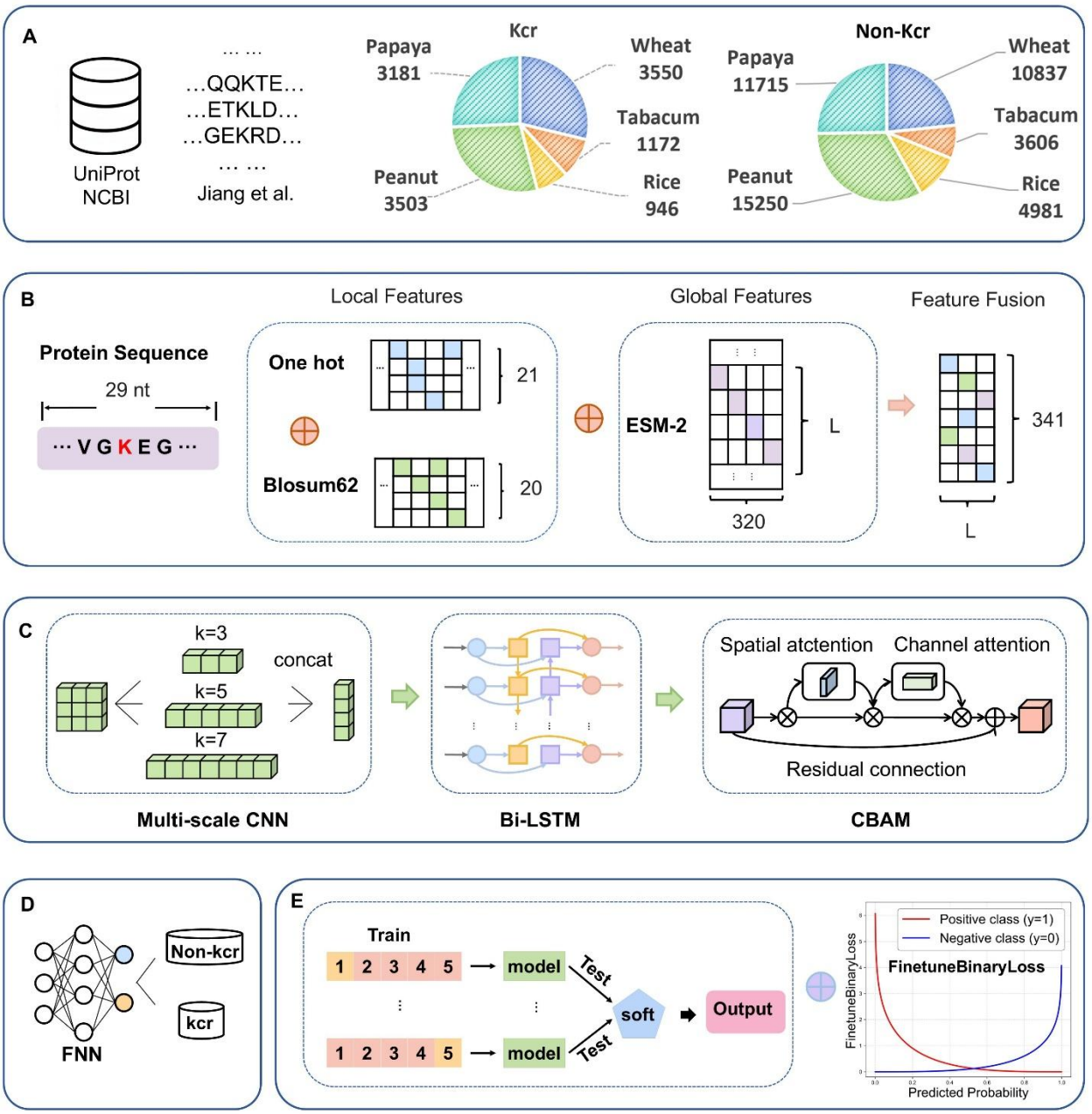


Fig. 1. Schematic of the MLC-KCR prediction framework.

(a) Multi-species Data Collection: Kcr modification data across five distinct species (papaya, wheat, tabacum, rice, and peanut) were collected to balance positive and negative samples for robust model training. (b) Hierarchical Feature Encoding: Local evolutionary and compositional information is encoded via One-hot and BLOSUM62, while pre-trained ESM-2 embeddings capture global contextual representations. These heterogeneous features are integrated into a fused feature matrix. (c) Hierarchical Deep Learning Architecture: A cascaded neural network is employed for hierarchical feature extraction, integrating Multi-scale CNN, BiLSTM, and CBAM. (d) Classification: Integrated deep features are passed through a feedforward neural network for classification. (e) Imbalance and Ensemble

Strategy: A cross-validation-based homogeneous ensemble strategy integrating the customized FinetuneBinaryLoss is employed, with the final output generated via soft voting.

2. Materials and Methods

2.1. Benchmark dataset

To train and test the model, we adopted the publicly available plant non-histone Kcr benchmark dataset constructed by Jiang et al [14]. In their dataset construction, non-histone Kcr sites were collected from five plant species (wheat, rice, tabacum, peanut, and papaya), and the corresponding protein sequences were retrieved from the UniProt [30] and NCBI [31] databases. Peptides of length 29 were extracted with the K (lysine) residue positioned at the center and 14 residues upstream and downstream. If a peptide had fewer than 14 residues on either side, the missing residues were replaced with X. Peptides whose central K residues were annotated as Kcr sites were designated as positive samples; otherwise, they were designated as negative samples. To eliminate redundancy and potential false negatives, the CD-HIT program [32] was applied at a sequence identity threshold of 40% during the dataset construction. Consequently, the resulting high-confidence dataset consisted of 12352 positive and 46389 negative samples.

To ensure a fair evaluation across multiple plant species and prevent data leakage, the total dataset was divided using a stratified strategy based on species. For each of the five plant species, the corresponding samples were randomly divided into two subsets at a 7:3 ratio, strictly preserving the original proportion of positive and negative samples. The larger subsets from all five species were merged to construct the global training set, totaling 41114 samples, consisting of 8644 positive and 32470 negative samples. The smaller subsets were merged to form the independent test set, comprising 17627 samples, with 3708 positive and 13919 negative. Crucially, this independent test set remained entirely unseen during model training and hyperparameter tuning, and was strictly reserved for the final stage to evaluate the model's true generalization capability. In this study, we directly adopted this predefined data split to ensure a fair comparison. The detailed sample distributions for each species are listed in Table 1.

Table 1. The distribution of positive and negative samples for each species in the benchmark dataset.

Species	Training Set		Independent Test Set	
	Positive	Negative	Positive	Negative
wheat	2484	7585	1066	3252
tabacum	820	2524	352	1082
rice	662	3486	284	1495

peanut	2452	10675	1051	4575
papaya	2226	8200	955	3515
Total samples	8644	32470	3708	13919

Values represent the number of sample sequences used for training and independent testing.

2.2. Feature coding schemes

Effective feature encoding is an essential step in translating raw protein sequences into numerical matrices suitable for deep learning models. To comprehensively capture the multifaceted characteristics of Kcr sites, we adopted three complementary encoding strategies: One-hot encoding for basic categorical representation, the BLOSUM62 matrix for evolutionary conservation, and the ESM-2 protein language model for deep contextual semantics.

2.2.1. One-Hot Encoding

One-hot encoding is a widely adopted method for representing categorical data in biological sequence analysis. It converts each amino acid into a unique, orthogonal binary vector, collectively forming a numerical feature matrix. For the Kcr site prediction task, we defined an alphabet of 21 characters, encompassing the 20 canonical amino acids and a placeholder 'X' used to pad sequences at the protein termini. The mapping rule is defined as follows:

$$\begin{gathered} A \rightarrow (1,0,0,\ldots,0) \\ C \rightarrow (0,1,0,\ldots,0) \\ \ldots \\ X \rightarrow (0,0,0,\ldots,1) \end{gathered} \tag{1}$$

In this representation, each amino acid is mapped to a 21-dimensional feature vector. Within this vector, only the index corresponding to the specific amino acid is activated (set to 1), while all other dimensions remain inactivated (set to 0). Applying this transformation position-by-position along a target sequence fragment of length $L = 29$ yields a 29×21 feature matrix. This straightforward scheme is computationally efficient, avoids the introduction of false ordinal relationships between amino acid types, and losslessly preserves sequence identity and precise positional information.

2.2.2. BLOSUM62 Encoding

While one-hot encoding preserves residue identity, it treats all amino acids as entirely orthogonal, disregarding the similar physicochemical properties and evolutionary substitution patterns they share during natural evolution. To address this limitation, we introduced the BLOSUM62 (BLOCKS Substitution Matrix) encoding [33]. Derived from highly conserved protein sequence alignments, this matrix quantifies the empirical probabilities of amino acid

substitutions over evolutionary history.

In this encoding, each natural amino acid is mapped to its corresponding row in the BLOSUM62 matrix, yielding a 20-dimensional numerical feature vector. For the placeholder 'X', a 20-dimensional zero vector is applied to maintain dimensional consistency. Thus, a target sequence fragment of length 29 is transformed into a 29×20 feature matrix. This evolution-based strategy assigns similar numerical weights to biochemically related amino acids, substantially assisting the model in recognizing functionally conserved motifs associated with Kcr modifications.

2.2.3. ESM-2 Encoding

Although the BLOSUM62 matrix captures evolutionary conservation, it is inherently static and context-independent. The feature vector for a specific amino acid remains identical regardless of its sequence position. To overcome this limitation and extract deeper biological semantics, we employed ESM-2, a large-scale, Transformer-based pre-trained model developed by Meta AI [34]. By undergoing self-supervised masked language modeling across massive protein databases, ESM-2 inherently learns the complex evolutionary rules and spatial structural properties embedded within protein sequences.

Unlike traditional static encodings, ESM-2 generates dynamic, context-aware representations. This means the embedding of an amino acid adapts to its local sequence microenvironment. Consequently, it implicitly encapsulates key biophysical properties such as secondary structure propensity, hydrophobicity, and local charge distribution. Since post-translational modifications like crotonylation are highly dependent on their local spatial and chemical environments, these dynamic features are highly valuable for accurate prediction. In this research, we utilized the esm2_t6_8M_UR50D variant to extract residue-level features, mapping each amino acid into a 320-dimensional dense vector. Consequently, for a sequence fragment of length 29, the ESM-2 encoding yields a 29×320 dense matrix, providing the model with a rich and biologically meaningful feature representation.

2.3. Model architecture

Predicting protein crotonylation sites requires capturing local physicochemical patterns and modeling long-range contextual dependencies. Additionally, dynamically filtering out background noise is essential for enhancing discriminative power. To address these challenges, we propose MLC-KCR, a synergistic deep learning framework. As illustrated in Figure 1, MLC-KCR employs a hierarchical strategy to process encoded feature matrices through

three complementary modules.

First, a Multi-scale Convolutional Neural Network utilizes parallel kernels of varying sizes to capture local structural motifs. This approach effectively mitigates the information loss commonly seen in single-scale CNNs. Next, a Bidirectional Long Short-Term Memory (BiLSTM) network models long-range sequential dependencies between distant amino acids, compensating for the local focus of convolutions. To further refine the extracted features, a Convolutional Block Attention Module (CBAM) adaptively recalibrates feature weights along both channel and spatial dimensions. This attention mechanism ensures that the network focuses on the most discriminative biological signals. Through this multi-level feature integration paradigm, our framework enables accurate and robust identification of potential Kcr sites.

2.3.1. Multi-scale Convolutional Neural Network

Biological sequences often contain functional motifs of varying lengths. Relying on a single convolution kernel size may fail to capture this structural diversity. To overcome this limitation, we designed a parallel Multi-scale CNN module. Given an input feature matrix $X \in R^{C_{in} \times L}$ (where $L = 29$ is the sequence length and C_{in} is the input feature dimension), the module employs three parallel 1D convolutional layers with distinct kernel sizes ($k \in \{3,5,7\}$). To ensure spatial alignment, a same padding strategy is applied to each layer. This guarantees a constant output sequence length across all branches. The output of the k-th convolutional branch, H_k, is computed as:

$$H_k = \mathrm{ReLU}\big(\mathrm{Conv1D}_k(X)\big) \tag{2}$$

The outputs from the three branches are concatenated along the channel dimension and regularized via a dropout layer. Finally, a 1×1 pointwise convolution is applied. This step compresses the concatenated high-dimensional features into a unified dimension C_{out}, thereby achieving cross-channel information fusion and yielding the local feature representation $H_{out} \in R^{C_{out} \times L}$.

2.3.2. Bidirectional Long Short-Term Memory

While the multi-scale CNN module excels at capturing local physicochemical patterns, it has a limited receptive field. Consequently, it struggles to model global sequential dependencies between distant amino acids. To address this, we integrated a two-layer Bidirectional Long Short-Term Memory (BiLSTM) network downstream of the convolutional module. At its core, the BiLSTM relies on LSTM units that employ internal gating mechanisms to

regulate information flow and mitigate the vanishing gradient problem [35,36].

In our framework, the BiLSTM processes the sequence features in two directions: a forward LSTM runs from the start to the end of the sequence ($x_1 \rightarrow x_T$), while a backward LSTM runs in reverse ($x_T \rightarrow x_1$). The final output for each position t is obtained by concatenating the forward hidden state $\overrightarrow{h_t}$ and the backward hidden state $\overleftarrow{h_t}$:

$$h_t^{bi} = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{3}$$

This bidirectional strategy ensures that the representation of each residue incorporates comprehensive global context from the entire sequence fragment, thereby complementing the local features extracted by the multi-scale CNN. Ultimately, the BiLSTM outputs a sequence of contextualized embeddings $H_{lstm} \in R^{2d_h \times L}$, where d_h is the hidden dimension of a single LSTM direction.

2.3.3. Convolutional Block Attention Module

Not all feature channels or sequence positions contribute equally to the identification of Kcr sites. To adaptively highlight informative features and suppress irrelevant noise, we integrated the Convolutional Block Attention Module (CBAM) [37]. The CBAM sequentially infers attention maps along two independent dimensions: channel and spatial. Given the intermediate feature map $H_{lstm} \in R^{2d_h \times L}$, the module first computes a channel attention map $M_c \in R^{2d_h \times 1}$. This is achieved by aggregating spatial information via global average-pooling and max-pooling, followed by a shared multi-layer perceptron (MLP). The channel attention map is formulated as:

$$M_c = \sigma\left(\mathrm{MLP}\left(\mathrm{AvgPool}(H_{lstm})\right) + \mathrm{MLP}\left(\mathrm{MaxPool}(H_{lstm})\right)\right) \tag{4}$$

where σ denotes the Sigmoid activation function. The input features are then element-wise multiplied by M_c to generate the channel-refined features, denoted as H'.

Next, to compute the spatial attention map $M_s \in R^{1 \times L}$, the module applies average-pooling and max-pooling along the channel axis. The pooled features are concatenated and convolved via a standard convolution. The spatial attention map is computed as:

$$M_s = \sigma\left(\mathrm{Conv}([\mathrm{AvgPool}(H'); \mathrm{MaxPool}(H')])\right) \tag{5}$$

where $[\cdot;\cdot]$ represents the concatenation operation. The final attention-enhanced feature map H_{attn} is obtained by multiplying the channel-refined features H' by the spatial attention weights:

$$H_{attn} = H' \odot M_s \tag{6}$$

where $\odot$ denotes element-wise multiplication. This dual-attention mechanism dynamically recalibrates the features, ensuring the network focuses on the most discriminative biological signals.

2.3.4. FinetuneBinaryLoss

Predicting protein modification sites faces a major challenge of inherent class imbalance. Non-modification sites outnumber true modification sites. Traditional Cross-Entropy loss is easily overwhelmed by the majority class, leading to poor sensitivity for the minority class. The standard Focal Loss mitigates this issue using a focusing parameter γ and a balancing weight α. the standard Focal Loss involves two hyperparameters, α and γ , which typically require a time-consuming joint grid search to find the optimal combination. This manual tuning process is not only computationally intensive but also results in a globally fixed α that fails to adapt to the dynamic fluctuations of class ratios within individual mini-batches.

To address this, we adopted the FinetuneBinaryLoss as implemented in the COMPASS-PTM framework [24] for our classification task. In this framework, the balancing factor α is dynamically computed based on the real-time class distribution, thereby eliminating the manual tuning of α. Consequently, the hyperparameter tuning process is simplified to adjusting only the focusing parameter γ, effectively reducing the overall parameter search space and the hyperparameter tuning burden.

For a given mini-batch with N_{pos} positive and N_{neg} negative samples, the adaptive weights α_p and α_n are inversely proportional to their frequencies:

$$\alpha_p = \frac{N_{neg} + \epsilon}{N_{pos} + N_{neg} + 2\epsilon} \tag{7}$$

$$\alpha_n = \frac{N_{pos} + \epsilon}{N_{pos} + N_{neg} + 2\epsilon} \tag{8}$$

where ϵ is a small smoothing factor (10^{-8}) to ensure numerical stability. Let $p \in [0,1]$ represent the predicted probability for the positive class, and $y \in \{0,1\}$ be the ground-truth label. The predicted probability for the true class, p_t, is defined as:

$$p_t = p \cdot y + (1 - p) \cdot (1 - y) \tag{9}$$

Correspondingly, its dynamic weight α_t is determined by:

$$\alpha_t = \alpha_p \cdot y + \alpha_n \cdot (1 - y) \tag{10}$$

The final FinetuneBinaryLoss is formulated as:

$$L_{FTB} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{11}$$

Integrating this dynamic adaptation into our framework offers two distinct advantages. First, it eliminates the computational burden of searching for an optimal global α. Second, it handles the high variance in class distribution

caused by random sampling. The dynamically computed α_t acts as a real-time stabilizer. It neutralizes local fluctuations and effectively smooths out gradient jitters during training.

2.3.5. Cross-Validation-Based Homogeneous Ensemble

The aforementioned mechanisms optimize the training process. However, a single deep learning model trained on a fixed data split often suffers from high variance. Data stochasticity and training noise cause this instability, which limits overall predictive performance**.** To address this, we implemented a homogeneous ensemble strategy based on five-fold cross-validation. This approach aggregates multiple predictions to smooth out local variations, effectively improving the robustness of the MLC-KCR framework [38,39].

As illustrated in **Figure 2**, all base models in this ensemble share the identical MLC-KCR architecture. We applied stratified sampling to divide the training set into five equal, mutually exclusive folds. This maintains the original positive-to-negative sample ratio in each fold and prevents bias from data distribution shifts. The training involves five independent rounds. In each round, four folds train a base model, and the remaining fold serves as validation. This yields five structurally identical models with distinct parameter spaces. During inference, unseen test sequences enter all five models simultaneously. The final output is derived via soft voting, by averaging the positive-class probabilities from all base models [40].

This homogeneous ensemble strategy provides two distinct advantages. First, by integrating collective knowledge learned from diverse data subsets, it effectively offsets the prediction variance of a single model and reduces the risk of local overfitting, thereby maximizing the model's generalization capability on unseen sequences. Second, when handling imbalanced data, aggregating multiple predictions prevents the final decision boundary from being dominated by the majority negative samples. This stabilizes the model's prediction confidence for the minority positive class, further mitigating the inherent bias induced by imbalanced datasets.

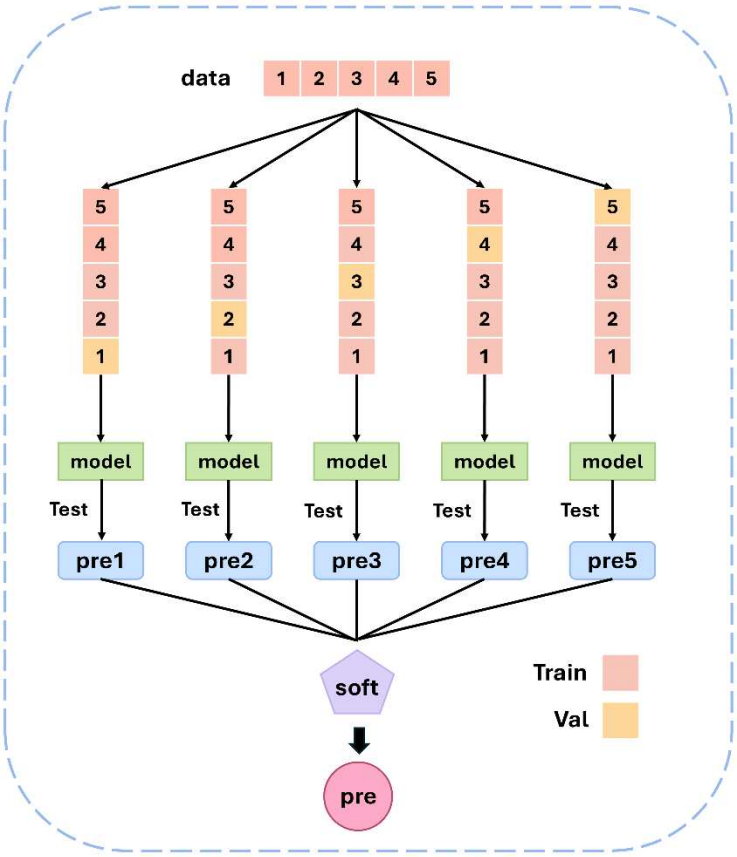


Fig. 2. Schematic of the five-fold cross-validation-based homogeneous ensemble strategy. The dataset is partitioned into five equal folds, where four folds are used for training, denoted as Train, and the remaining one for validation, denoted as Val, in each round. Final predictions are generated via soft voting.

2.4. Performance evaluation

The predictive performance of MLC-KCR was systematically assessed using six standard classification metrics: sensitivity (Sn), specificity (Sp), accuracy (Acc), area under the ROC curve (AUC), Matthews correlation coefficient (MCC), and F1-score (F1). The mathematical definitions of these metrics are provided below:

$$\begin{cases} \mathrm{Sn} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ \mathrm{Sp} = \dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \\ \mathrm{MCC} = \dfrac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \\ \mathrm{Acc} = \dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \\ \mathrm{AUC} = \displaystyle\int_0^1 \mathrm{Sn}\cdot \mathrm{d}(1-\mathrm{Sp}) \\ \mathrm{F1} = \dfrac{2\times\mathrm{TP}}{2\times\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \end{cases} \tag{12}$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.

Inherent class imbalances in biological sequence datasets often render standard accuracy misleading or overly optimistic. Consequently, MCC, AUC, and F1 served as the primary benchmarks for performance evaluation, as they provide a more robust assessment of the model's capability to identify the minority positive class. While MCC provides a balanced evaluation by integrating all four facets of the confusion matrix, AUC measures the overall discriminative capability, and F1 reflects the harmonic mean of precision and recall. Meanwhile, Sn, Sp, and Acc were reported as general reference metrics [41]. Under established statistical frameworks, higher values for these metrics correspond to better discriminative capability and enhanced model reliability.

3. Results

3.1. Classification performance evaluation of MLC-KCR on the benchmark dataset

To evaluate the performance of MLC-KCR for plant Kcr site prediction, we employed five-fold cross-validation on the global training set and an independent test set. For the independent test set, the model's performance was assessed based on a single evaluation of the soft-voting ensemble model to demonstrate its predictive potential. The predictive performance was assessed using six standard metrics, including Sn, Sp, Acc, MCC, F1-score, and AUC.

As summarized in Table 2, during the five-fold cross-validation on the global training dataset, the model exhibited consistent and balanced predictive performance across all evaluation metrics. Specifically, the comparable Sn and Sp values indicate balanced classification behavior under imbalanced data conditions. Furthermore, all evaluation metrics exhibited low standard deviations across different data partitions, with MCC and F1-score particularly remaining at stable levels, reflecting reliable overall predictive capability. As illustrated in Figure 3A, the ROC curves from different folds were closely aligned, with minimal variation in AUC values, further confirming the stability and robustness of the training framework. Subsequently, the independent test set was used to evaluate the generalization ability of the model. As shown in Figure 3B, MLC-KCR achieved an AUC of 0.911 on unseen data. Notably, this performance was slightly higher than the mean AUC obtained from cross-validation, suggesting robust generalizability and stable predictive performance on unseen data.

Table 2. Five-fold cross-validation results of MLC-KCR on each fold of the global training set.

Fold	Sn	Sp	Acc	AUC	MCC	F1
0	0.812	0.833	0.828	0.895	0.571	0.665
1	0.852	0.810	0.822	0.904	0.581	0.669
2	0.794	0.847	0.837	0.898	0.581	0.673

3	0.837	0.814	0.817	0.900	0.562	0.656
4	0.802	0.838	0.829	0.894	0.564	0.660
Mean ± Std	0.819±0.024	0.828±0.016	0.827±0.007	0.898±0.004	0.572±0.009	0.665±0.007

Results for Folds 0–4 represent the performance of each individual fold, while the final row displays the mean ± standard deviation across all five folds.

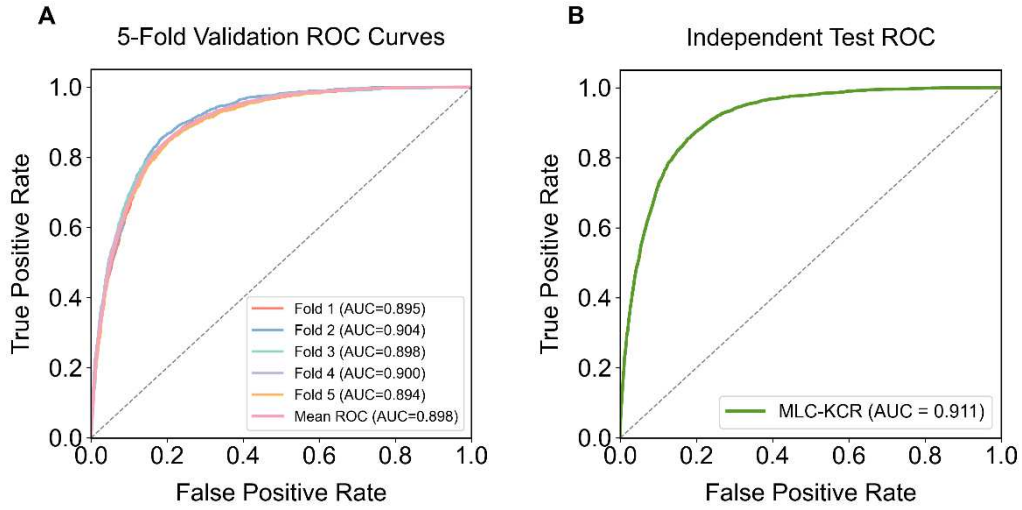


Fig. 3. ROC curve analysis illustrating the predictive performance of MLC-KCR. (A) Individual and mean ROC curves obtained from 5-fold cross-validation on the integrated dataset, yielding an average AUC of 0.898. (B) Performance of the MLC-KCR model on the independent test set, which achieved a robust AUC value of 0.911.

3.2. Comparison of different coding methods

This section compares several sequence encoding strategies, including One-hot, BLOSUM62, ESM-2, Z-scale, and AAindex. Considering the class imbalance in the dataset, MCC and AUC were selected as the primary evaluation metrics, while Sn and Sp were used to assess class-specific predictive capabilities. The results of the five-fold cross-validation are presented in Figure 4. Among the individual encoding methods, ESM-2 achieved the highest Sp, indicating that embeddings from protein language models can effectively capture contextual sequence information. However, its Sn was lower than that of certain hand-crafted features, suggesting that relying solely on global contextual representations is insufficient for identifying all positive Kcr sites. In contrast, BLOSUM62 achieved a relatively high Sn, demonstrating its efficacy in recognizing Kcr-related sequence patterns; however, its MCC was

lower compared to the fusion strategy, indicating inherent limitations in its overall discriminative capacity. Meanwhile, Z-scale and AAindex individually yielded moderate performance, further highlighting the limitations of single encoding schemes.

To overcome the limitations of individual feature representations, we adopted a feature fusion strategy combining One-hot, BLOSUM62, and ESM-2. This configuration achieved the best overall performance among all evaluated encoding strategies, yielding a mean MCC of 0.572 ± 0.009 and a mean AUC of 0.898 ± 0.004. This performance advantage stems primarily from the complementarity of the three encoding schemes. Specifically, One-hot encoding preserves precise positional features, BLOSUM62 introduces evolutionary substitution information, and ESM-2 embeddings provide contextual sequence representations. By integrating these heterogeneous features, the model leverages both local sequence patterns and global contextual information, thereby attaining robust discriminative capability. Therefore, the tripartite fusion of One-hot, BLOSUM62, and ESM-2 was adopted as the final sequence representation for MLC-KCR.

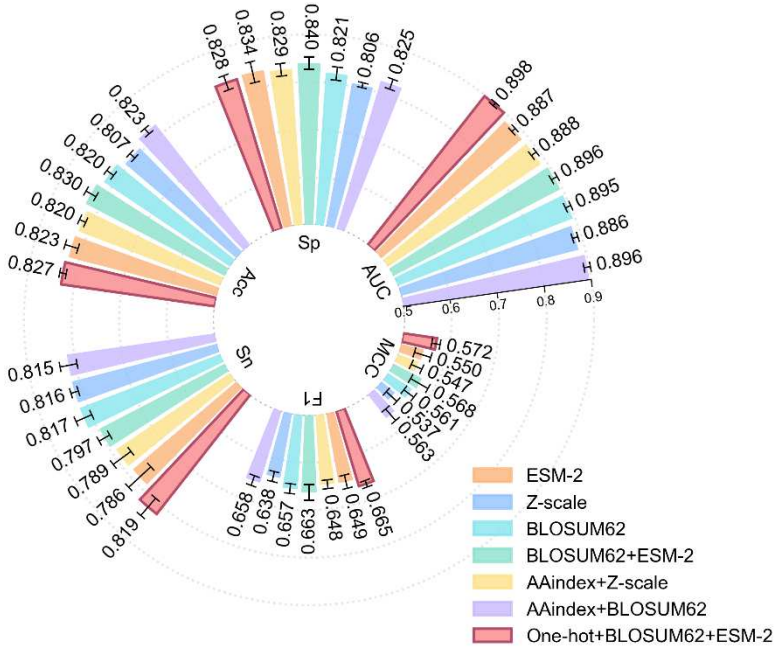


Fig. 4. Performance of different feature encoding schemes in 5-fold cross-validation on the global training set. Six key metrics are presented in distinct radial sectors, with bar heights and error bars representing the mean scores and standard deviations, respectively. The onehot+blosum62+esm2 scheme (red outline) achieved the highest overall performance across major metrics, demonstrating robust predictive capability and stability.

3.3. Ablation experiments for MLC-KCR

To evaluate the contribution of each submodule to the overall architecture, we conducted a systematic ablation study. The results from five-fold cross-validation are summarized in Figure 5. Among the standalone modules, BiLSTM exhibited the most robust baseline performance, achieving a mean MCC of 0.549 ± 0.026 and an AUC of 0.894 ± 0.007. This result indicates that capturing long-range bidirectional dependencies is essential for extracting discriminative features from sequence representations. The CBAM-only model yielded a much lower MCC of 0.467 ± 0.016, suggesting that attention mechanisms alone are insufficient for effective feature extraction without proper sequence modeling. When integrating the Multi-scale CNN (MS-CNN) with BiLSTM, the performance effectively improved, with the mean MCC increasing to 0.562 ± 0.019 and Sp reaching 0.833 ± 0.020. This suggests that combining local motifs captured by MS-CNN with global sequential context modeled by BiLSTM enhances the model's ability to capture complex sequence patterns.

The full MLC-KCR model achieved the best overall performance with a well-balanced metric profile: MCC of 0.572 ± 0.009, AUC of 0.898 ± 0.004, Acc of 0.827 ± 0.007, and F1 of 0.665 ± 0.007. The higher MCC and F1 values demonstrate the advantage of the proposed three-stage "extraction–modeling–refinement" architecture. Specifically, after local motifs and global context are integrated by MS-CNN and BiLSTM, the CBAM module further applies adaptive channel-wise and spatial weighting. This attention mechanism effectively suppresses redundant information and improves the quality of feature representations. Overall, these results validate that each architectural component contributes uniquely to the final performance, supporting the effectiveness of the integrated design of MLC-KCR.

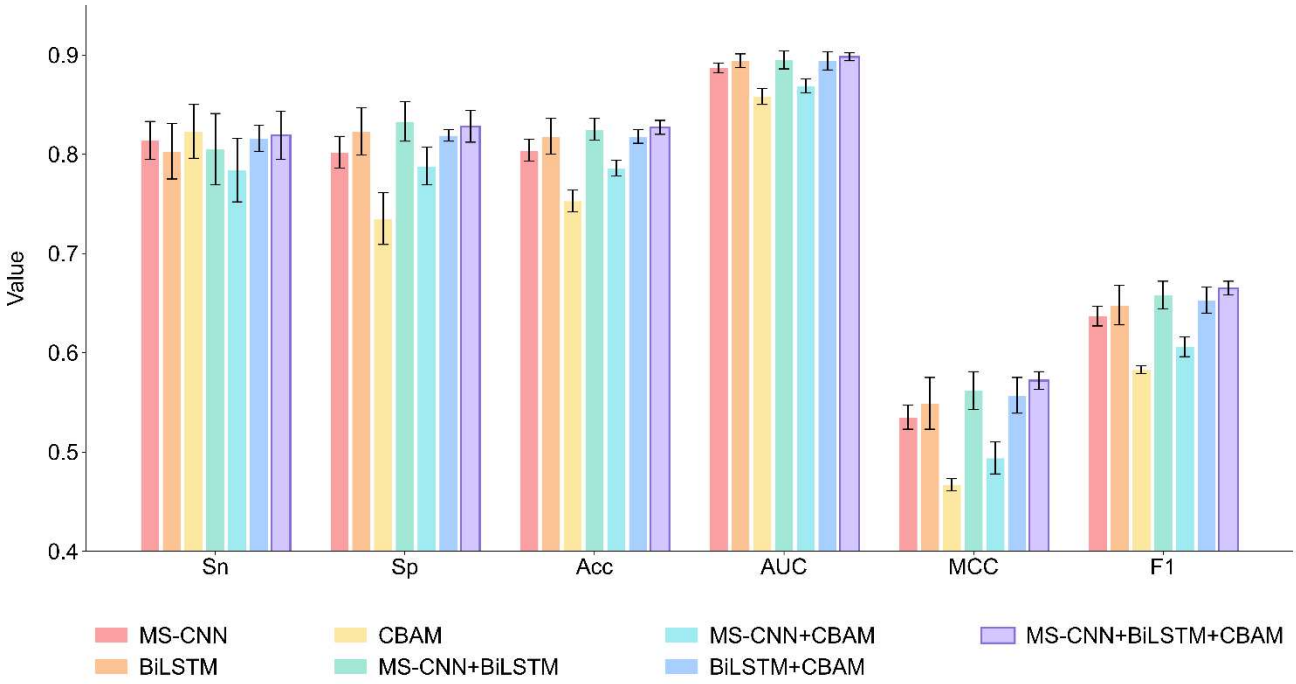


Fig. 5. Ablation results of MLC-KCR based on 5-fold cross-validation on the global training set. The bar chart displays the average performance of seven model combinations, with error bars representing the standard deviation. The MS-CNN+BiLSTM+CBAM model (purple bars) excels in key metrics such as AUC, MCC, and F1, thereby validating the effectiveness and rationality of our proposed three-module architecture design.

3.4. Comparison of Imbalance Strategies

Protein modification site prediction is often affected by notable class imbalance, which can bias models toward the majority class and weaken their ability to identify positive samples. To systematically evaluate different imbalance-handling strategies, we compared Downsampling, Focal Loss, and FinetuneBinaryLoss under both single-model and ensemble learning frameworks. The results presented in this section are based on a single independent test evaluation for each optimal model configuration within the respective strategies. The results are summarized in Figure 6.

In the single-model experiments, Downsampling increased Sn by randomly removing negative samples. However, this static preprocessing inevitably causes the permanent loss of key feature information hidden within the discarded majority-class samples, leading to a notable decrease in Sp. This indicates that simply discarding majority-class samples makes the model overly sensitive to positive samples at the cost of a higher false-positive rate, thereby

constraining overall metrics such as MCC and F1-score. Focal Loss alleviated the class imbalance to some extent by focusing on hard-to-classify samples; however, its globally fixed weighting mechanism fails to adapt to batch-level fluctuations, resulting in suboptimal performance in metrics such as MCC and F1-score. To address these limitations, we employed FinetuneBinaryLoss, which adaptively adjusts sample weights based on the label distribution within each mini-batch. This enables the model to enhance the learning capacity for minority-class samples while maintaining stable modeling of the majority class, thereby achieving the most balanced performance profile among all single-model strategies.

To further improve model robustness, a homogeneous ensemble learning strategy was introduced. By integrating multiple independently trained models, ensemble learning reduces prediction variance and improves the model's generalization ability across different data partitions. This strategy exhibits strong complementarity with FinetuneBinaryLoss: the former enhances prediction stability through model-level integration, while the latter mitigates the bias induced by class imbalance through optimization-level adaptive reweighting. Consequently, their combination achieved the best overall performance among all evaluated methods. Meanwhile, compared with Focal Loss, FinetuneBinaryLoss reduces the reliance on manually setting the class-balancing parameter α and eliminates the need for exhaustive joint parameter tuning. Therefore, only the focusing parameter γ needs to be adjusted, which effectively reduces the hyperparameter tuning burden in imbalanced learning scenarios. Therefore, this study adopts the combination of FinetuneBinaryLoss and homogeneous ensemble learning for MLC-KCR, providing a robust solution for imbalanced biological sequence prediction.

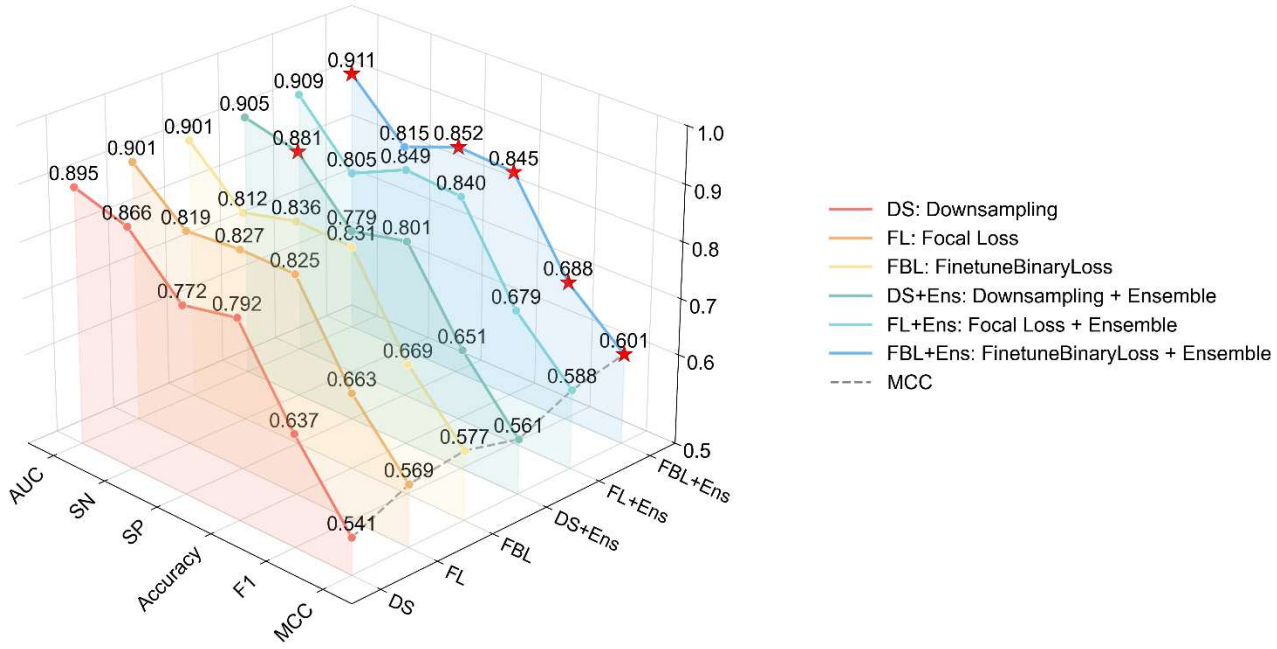


Fig. 6. Multidimensional visualization of performance under different class imbalance handling strategies. This three-dimensional plot illustrates the performance profiles of six strategies across multiple evaluation metrics. The x-axis represents the strategy categories, the y-axis denotes the six core performance metrics, and the z-axis indicates the corresponding values. Colored curves connect the scores of each strategy across metrics, red asterisks highlight the best values for each metric, and the gray dashed line depicts the variation of MCC across strategies. The results demonstrate that the FBL+Ens strategy (blue curve) achieves robust overall classification performance in key metrics, validating the effectiveness of this combined strategy.

3.5. Performance comparison with existing methods

To comprehensively evaluate the performance of MLC-KCR and systematically assess its effectiveness, we conducted comparative experiments against current state-of-the-art predictors. First, we assessed model robustness using rigorous five-fold cross-validation on the training set and compared it with the leading method, PlantNh-Kcr. Second, we comprehensively benchmarked its performance against five existing methods on an independent test set. To ensure fairness, we strictly followed the evaluation protocol of PlantNh-Kcr, repeating the independent test experiments 10 times under different random seeds and reporting the averaged results.

In the cross-validation, MLC-KCR demonstrated robust overall predictive performance and stability. As shown in Figure 7, its AUC reached 89.8% and its MCC reached 57.2%. Although its Sn of 81.9% was slightly lower than the 82.1% of PlantNh-Kcr, its Sp improved from 81.0% to 82.8%. Furthermore, its MCC increased from 55.1% to

57.2%. In imbalanced modification site prediction, a trade-off between Sn and Sp is a common phenomenon. A higher Sn is often accompanied by a predictive bias toward the minority class. This bias can reduce the ability to discriminate negative samples. In contrast, MLC-KCR mitigates this bias by combining an adaptive reweighting mechanism with a homogeneous ensemble strategy. This enables the model to achieve more balanced positive and negative sample recognition while improving overall discriminative ability. Consequently, this manifests as a slight decrease in Sn alongside an increase in Sp. The improvement in MCC confirms that this trade-off effectively enhances the overall classification performance, further supporting the effectiveness of the model in class-imbalanced scenarios.

The results on the independent test set are presented in Table 3. MLC-KCR outperformed other methods across most evaluation metrics. Some models, such as nhKcr and DeepKcrot, achieved higher Sn values of 87.6% and 82.8%, respectively. However, their Sp decreased to 77.3% and 77.9%, indicating a deficiency in negative sample recognition. In contrast, MLC-KCR maintained an Sn of 81.9% while achieving the highest Sp of 84.8%, reflecting a more balanced classification capability. Furthermore, its MCC of 59.8%, AUC of 91.0%, and F1-score of 68.6% all ranked first. These superior metrics further indicate that the model achieves a better balance between precision and recall. Notably, we compared MLC-KCR with the next-best model, PlantNh-Kcr, in the independent test. The Sn of MLC-KCR remained closely consistent, while its Sp was superior. Additionally, its MCC increased by 2.6 percentage points. These metrics demonstrate robust generalization consistency. This result suggests that the adaptive feature learning and ensemble strategies can effectively capture more stable sequence patterns. Consequently, this enhances the overall reliability of the model on unseen data.

Overall, these results demonstrate that MLC-KCR achieves a favorable balance between Sn and Sp, resulting in robust and reliable predictive performance. These characteristics establish MLC-KCR as a practical and efficient tool in the field of protein Kcr site prediction.

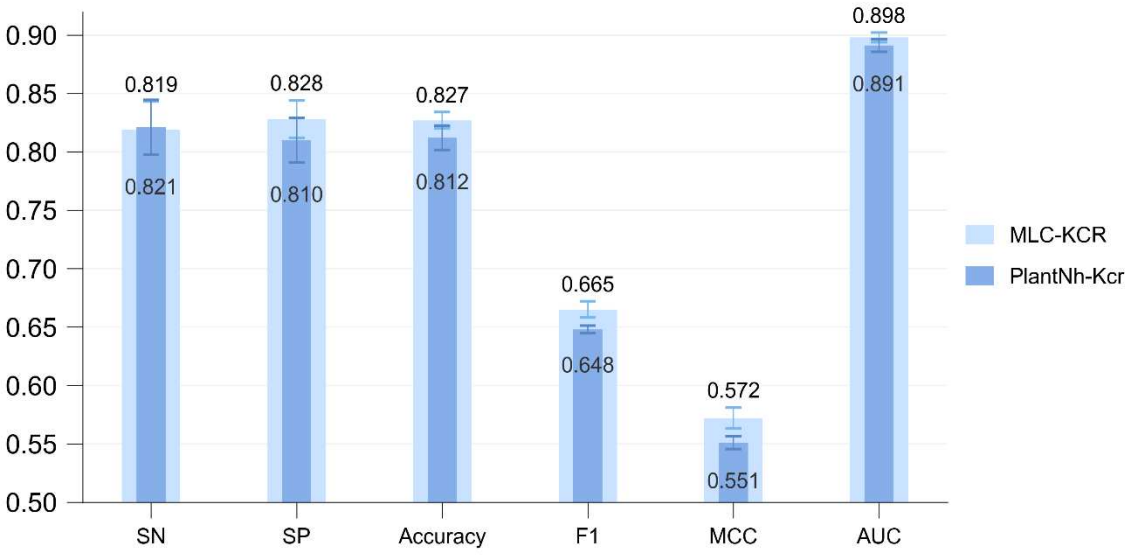


Fig. 7. Performance comparison of MLC-KCR and PlantNh-Kcr in 5-fold cross-validation. This bar chart compares the 5-fold cross-validation results between MLC-KCR and PlantNh-Kcr across six evaluation metrics. The height of each bar represents the mean performance value, with error bars indicating the standard deviation. Overall, MLC-KCR outperforms PlantNh-Kcr in key metrics such as MCC and AUC, demonstrating robust overall classification performance.

Table 3. Performance comparison of MLC-KCR and five existing methods on the independent test set.

Models	Sn(%)	Sp(%)	Acc (%)	F1(%)	MCC(%)	AUC(%)
iKcr_CNN	77.2±1.07	82.0±0.79	81.0±0.45	63.1±0.39	52.5±0.51	87.6±0.18
DeepKcrot	82.8±1.53	77.9±1.46	78.9±0.85	62.3±0.56	51.9±0.59	87.6±0.26
nhKcr	**87.6±2.42**	77.3±3.26	79.4±3.26	64.3±1.61	55.1±1.63	89.0±0.92
CapsNh-Kcr	76.4±3.59	84.3±1.59	83.1±0.71	65.5±0.47	55.6±0.58	89.2±0.09
PlantNh-Kcr	81.1±3.23	83.3±2.09	82.8±0.99	66.5±0.50	57.2±0.50	89.9±0.19
MLC-KCR	81.9±0.95	**84.8±0.63**	**84.2±0.32**	**68.6±0.24**	**59.8±0.28**	**91.0±0.04**

All values are presented as mean ± standard deviation. The best-performing results are highlighted in bold.

3.6. Comparison of Species-specific and Multi-species Training for Plant Kcr Site Prediction

To evaluate the effectiveness of the multi-species joint training strategy, we compared species-specific training with multi-species joint training. In the species-specific setting, the model was trained and tested using data from a single target species. In the multi-species setting, the model was trained using data from all species and evaluated separately on each species-specific test set. The results are summarized in Figure 8.

Overall, multi-species training consistently outperformed species-specific training across all tested species.

Compared with species-specific models, the multi-species model achieved higher MCC and AUC values for all species, indicating that integrating data from different species improves both discriminative ability and generalization performance. The largest improvements were observed in tabacum and rice, where MCC increased by 0.145 and 0.090, respectively, and AUC increased by 0.067 and 0.054, respectively. Peanut and papaya also benefited from multi-species training, although the gains were smaller. In the rice dataset, multi-species training increased Sn from 0.718 to 0.894, while Sp decreased slightly from 0.836 to 0.806, indicating a stronger ability to identify positive samples, a common trade-off in class-imbalanced scenarios.

Compared with species-specific models, the unified model trained on multi-species data achieved better performance across all plant species test sets. These results suggest that Kcr-related features are shared to some extent among different plant species, and that joint training can effectively learn these common patterns. Therefore, multi-species joint training provides a more effective alternative to species-specific models for accurate Kcr site prediction. A single model trained on multi-species data can provide reliable predictions across different plant species while maintaining stable performance. As a result, users can directly apply the model to protein sequences from different plant species without selecting species-specific predictors or switching between different models in advance. This simplifies the prediction workflow, improves the generality and scalability of the method, and provides a basis for the development of a general Kcr site prediction web server.

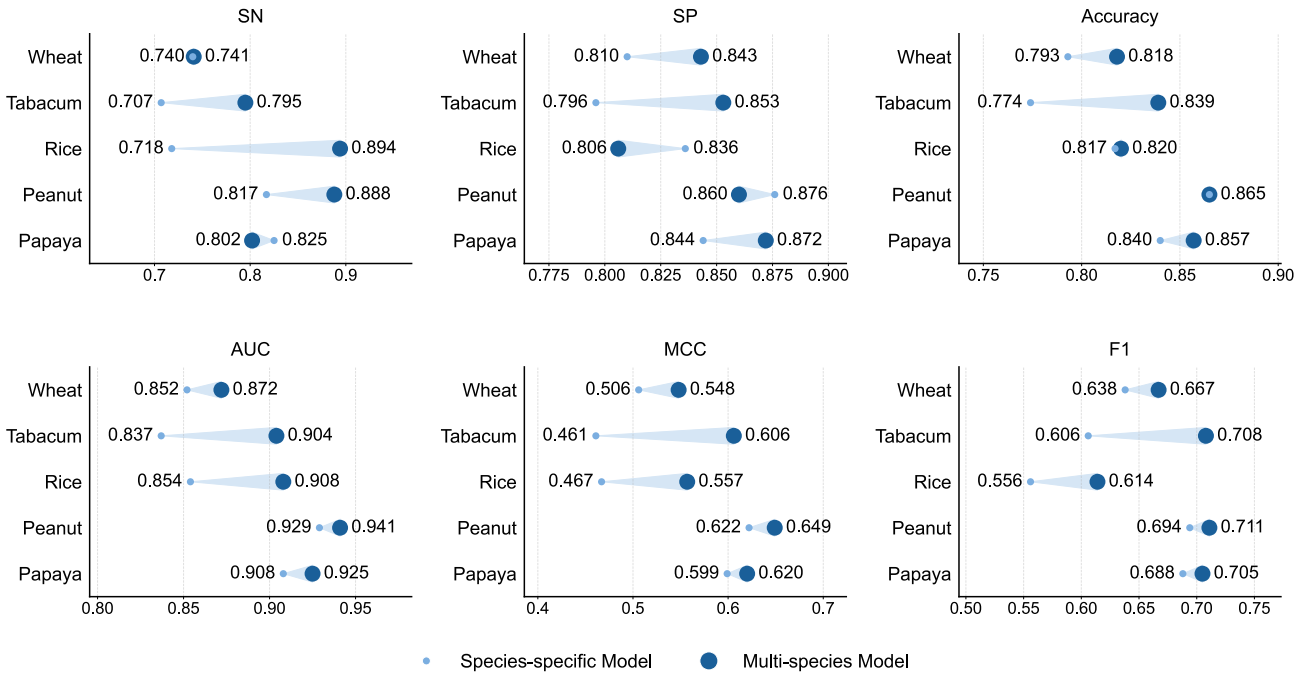


Fig. 8. Performance comparison of species-specific and multi-species training on the independent test set. This plot compares the performance of the two training strategies across six evaluation metrics, with light dots representing the species-specific model and dark dots representing the multi-species model. The results demonstrate that the multi-species model outperforms the species-specific model in key metrics such as AUC and MCC, validating the effectiveness of the joint training strategy and indicating robust generalization capability across multiple plant species.

3.7. Motif Analysis of Lysine Crotonylation Sites

To systematically characterize the sequence features surrounding lysine crotonylation sites, the kpLogo [42] tool was employed to compare the flanking sequences of Kcr and non-Kcr sites. This study selected a sequence window centered on the modified lysine, comprising 14 residues upstream and 14 downstream (29 residues in total) for analysis. As illustrated in Figure 9, the kpLogo results show an asymmetric distribution of residues around the modification sites.

Upstream of the Kcr site, a continuous enrichment of lysine (K) spans positions 1 to 13, peaking between positions 6 and 13, establishing an extended basic microenvironment. The immediate flanking regions show specific structural and chemical patterns. Position 14 exhibits an enrichment of the acidic residue glutamic acid (E) alongside aromatic residues tyrosine (Y) and phenylalanine (F). Position 16 is similarly enriched with acidic residues, primarily

glutamic acid (E) and aspartic acid (D). This indicates that the Kcr site is flanked by negative charges and hydrophobic aromatic groups. Further downstream from positions 20 to 24, lysine (K) is again enriched. In contrast, serine (S) and proline (P) are broadly depleted across the flanking regions, suggesting their presence may hinder the crotonylation process structurally or chemically.

Overall, the upstream poly-K cluster, the adjacent acidic and aromatic residues, and the depletion of serine and proline constitute a specific local sequence context for Kcr. These distinctive motif features provide a biological basis for employing the advanced feature extraction modules in the MLC-KCR model, which are designed to capture such complex contextual dependencies and ensure accurate site-specific prediction.



Fig. 9. kpLogo analysis of Kcr sites. The plot displays position-specific residue enrichment above the axis and depletion below the axis relative to the central crotonylation site at position 15. Symbol height indicates statistical significance measured by the negative log10 P-value.

3.8. A User-Friendly Web Server for MLC-KCR

To maximize the utility and accessibility of the proposed MLC-KCR model, a user-friendly web server was developed specifically for the identification of Kcr sites. Although deep learning approaches exhibit strong predictive performance, their deployment typically requires specialized programming expertise and the configuration of complex computational environments. This poses a practical technical barrier for many experimental biologists. The MLC-KCR online platform bridges this gap by providing an intuitive interface, enabling researchers to focus on biological validation rather than computational implementation.

As illustrated in Figure 10, the platform offers a streamlined and efficient workflow. Users can submit their target protein sequences by either pasting them directly into the input text box in FASTA format or uploading a FASTA file for high-throughput batch analysis. Importantly, the platform does not require users to specify the plant species, as it automatically performs predictions across different plant species using the multi-species trained model. Upon clicking the submit button, the server utilizes the pre-trained MLC-KCR model to rapidly process the input data. The prediction results detailing the potential Kcr sites are then presented directly on the web interface. For user convenience during large-scale data analysis, the platform also supports returning the detailed prediction results via email. The MLC-KCR web server is freely accessible to the public at https://unblock-research-reshuffle.ngrok-free.dev.

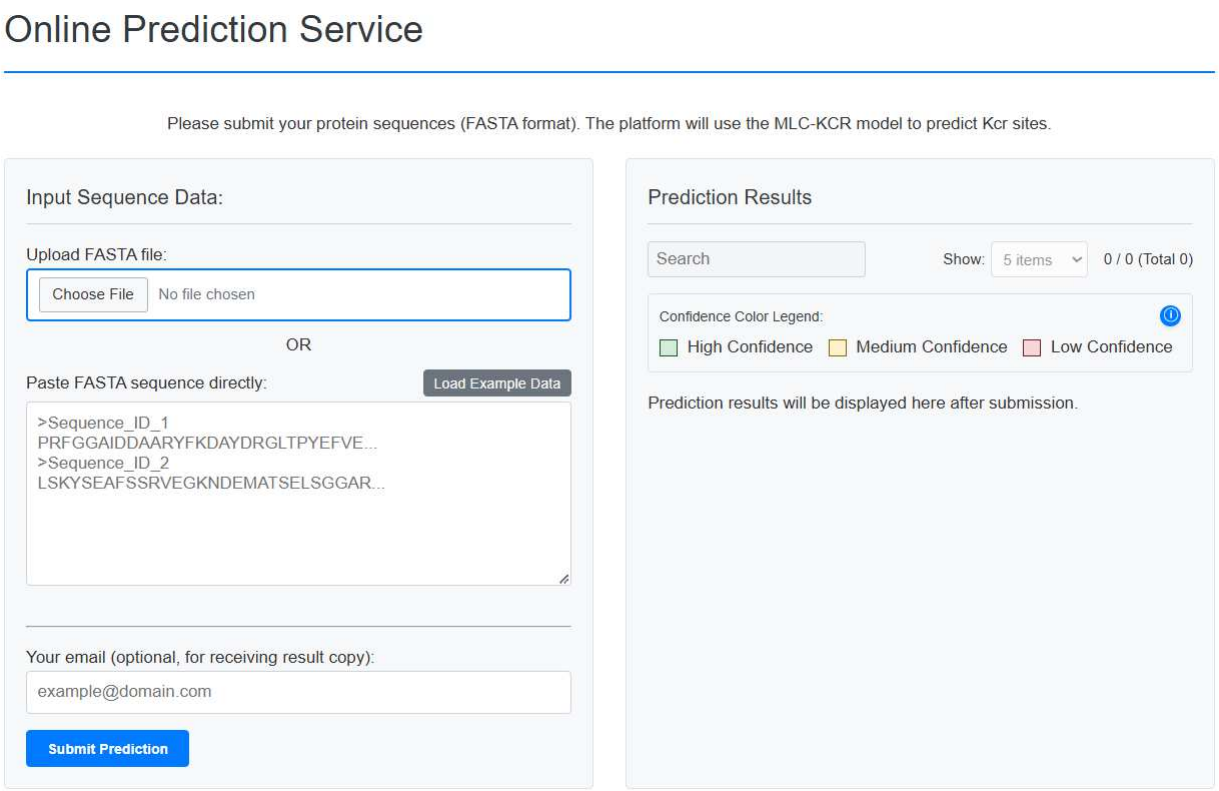


Fig. 10. Homepage of the MLC-KCR web server. The interface allows users to submit protein sequences in FASTA format via file upload or text paste and displays the prediction results.

4. Conclusions

This study proposes the MLC-KCR framework for multi-species plant Kcr site prediction, aiming to enhance

the predictive performance and generalization ability in identifying Kcr sites. The model fuses One-hot, BLOSUM62, and ESM-2 encodings to construct sequence features from the perspectives of local sequence characteristics, evolutionary substitution information, and global semantic representations, respectively, thereby obtaining a more comprehensive representation. Structurally, MLC-KCR employs a cascaded architecture consisting of Multi-scale CNN, BiLSTM, and CBAM. The Multi-scale CNN extracts multi-scale local motif features, the BiLSTM models bidirectional long-range dependencies of the sequence, and the CBAM further optimizes feature representations through channel and spatial attention mechanisms. Ablation experiments demonstrate the complementary contributions of these modules, collaboratively improving the predictive capacity of the model. To address the class imbalance issue in Kcr prediction, this study integrates FinetuneBinaryLoss with a homogeneous ensemble learning strategy. These two strategies complement each other at the optimization and model levels, effectively enhancing the learning capacity for minority-class samples and overall predictive stability. In multi-species evaluations, a single model trained on multi-species data outperformed species-specific models across all individual plant test sets, indicating the existence of shared Kcr modification patterns among different plant species. Consequently, MLC-KCR can provide stable and reliable predictions across diverse plant species without the need to build separate models for specific species. Finally, to enhance the practical application value of our method, we developed an online prediction platform based on MLC-KCR. Importantly, the platform does not require users to specify the plant species, allowing researchers to directly submit protein sequences and obtain accurate Kcr site predictions via a web interface.

Although MLC-KCR has achieved promising performance in plant protein Kcr site prediction, several limitations remain. First, the model primarily relies on primary sequence information and lacks explicit three-dimensional structural features, which may limit its capacity to characterize spatially dependent modification sites. Second, without incorporating biophysical information at the structural level, the model might not fully capture certain modification mechanisms driven by long-range spatial interactions. Furthermore, despite the implementation

of an adaptive loss function and ensemble learning strategies, the recognition ability for minority-class samples could still be compromised under severe class imbalance scenarios. Future research could address these limitations by integrating protein structural information, such as structures predicted by AlphaFold2, and employing graph neural networks or three-dimensional convolutional networks to more accurately model the spatial relationships between residues. Additionally, exploring self-supervised learning and large-scale pre-training strategies on unlabeled sequences could further enhance the model's generalization capability for low-resource species and diverse protein families.

List of abbreviations

Kcr: Lysine crotonylation

ESM-2: Evolutionary Scale Modeling 2

CNN: Convolutional Neural Network

BiLSTM: Bidirectional Long Short-Term Memory

CBAM: Convolutional Block Attention Module

Sn: Sensitivity

Sp: Specificity

Acc: Accuracy

MCC: Matthews Correlation Coefficient

F1: F1-score

AUC: Area Under the Receiver Operating Characteristic Curve

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The source code and datasets generated in this study are publicly available at GitHub: https://github.com/linlinhuihui/MLC-KCR.

Competing interests

The authors declare that they have no competing interests

Funding

This work was partially supported by the Scientific Research Plan of the Department of Education of Jiangxi Province, China (GJJ2400909, GJJ2402711). These funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Authors' contributions

HZ: Formal analysis, Methodology, Visualization, Writing – Original Draft, Writing – Review & Editing. JJ: Conceptualization, Resources, Supervision, Writing – Review & Editing. HH: Data Curation, Investigation, Validation, Writing – Review & Editing. All authors read and approved the final manuscript.

Acknowledgements

The authors are grateful for the constructive comments and suggestions made by the reviewers.

References and Notes

1. Suskiewicz MJ (2024) The logic of protein post-translational modifications (PTMs): Chemistry, mechanisms and evolution of protein regulation through covalent attachments. Bioessays 46:e2300178 https://doi.org/10.1002/bies.202300178
2. Dakup PP, Feng S, Shi T, et al. (2023) Targeted Quantification of Protein Phosphorylation and Its Contributions towards Mathematical Modeling of Signaling Pathways. Molecules 28 https://doi.org/10.3390/molecules28031143
3. Wu X, Xu M, Geng M, et al. (2023) Targeting protein modifications in metabolic diseases: molecular mechanisms and targeted therapies. Signal Transduct Target Ther 8:220 https://doi.org/10.1038/s41392-023-01439-y

4. Liu R, Wu J, Guo H, et al. (2023) Post-translational modifications of histones: Mechanisms, biological functions, and therapeutic targets. MedComm (2020) 4:e292 https://doi.org/10.1002/mco2.292
5. Park J, Lee K, Kim K, et al. (2022) The role of histone modifications: from neurodevelopment to neurodiseases. Signal Transduct Target Ther 7:217 https://doi.org/10.1038/s41392-022-01078-9
6. Zhao H, Han Y, Zhou P, et al. (2024) Protein lysine crotonylation in cellular processions and disease associations. Genes Dis 11:101060 https://doi.org/10.1016/j.gendis.2023.06.029
7. Xie JY, Ju J, Zhou P, et al. (2024) The mechanisms, regulations, and functions of histone lysine crotonylation. Cell Death Discov 10:66 https://doi.org/10.1038/s41420-024-01830-w
8. Tan M, Luo H, Lee S, et al. (2011) Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell 146:1016 https://doi.org/10.1016/j.cell.2011.08.008
9. Liu S, Xue C, Fang Y, et al. (2018) Global Involvement of Lysine Crotonylation in Protein Modification and Transcription Regulation in Rice. Mol Cell Proteomics 17:1922 https://doi.org/10.1074/mcp.RA118.000640
10. Zhang N, Wang S, Zhao S, et al. (2023) Global crotonylatome and GWAS revealed a TaSRT1-TaPGK model regulating wheat cold tolerance through mediating pyruvate. Sci Adv 9:eadg1012 https://doi.org/10.1126/sciadv.adg1012
11. Sun H, Liu X, Li F, et al. (2017) First comprehensive proteome analysis of lysine crotonylation in seedling leaves of Nicotiana tabacum. Sci Rep 7:3013 https://doi.org/10.1038/s41598-017-03369-6
12. Liu K, Yuan C, Li H, et al. (2018) A qualitative proteome-wide lysine crotonylation profiling of papaya (Carica papaya L.). Sci Rep 8:8230 https://doi.org/10.1038/s41598-018-26676-y
13. Xu M, Luo J, Li Y, et al. (2021) First comprehensive proteomics analysis of lysine crotonylation in leaves of peanut (Arachis hypogaea L.). Proteomics 21:e2000156 https://doi.org/10.1002/pmic.202000156
14. Jiang Y, Yan R, Wang X (2024) PlantNh-Kcr: a deep learning model for predicting non-histone crotonylation sites in plants. Plant Methods 20:28 https://doi.org/10.1186/s13007-024-01157-8
15. Chen YZ, Wang ZZ, Wang Y, et al. (2021) nhKcr: a new bioinformatics tool for predicting crotonylation sites on human nonhistone proteins based on deep learning. Brief Bioinform 22 https://doi.org/10.1093/bib/bbab146
16. Ju Z, He JJ (2017) Prediction of lysine crotonylation sites by incorporating the composition of k-spaced amino acid pairs into Chou's general PseAAC. J Mol Graph Model 77:200 https://doi.org/10.1016/j.jmgm.2017.08.020
17. Qiu WR, Sun BQ, Xiao X, et al. (2018) iKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. Genomics 110:239 https://doi.org/10.1016/j.ygeno.2017.10.008
18. Liu Y, Yu Z, Chen C, et al. (2020) Prediction of protein crotonylation sites through LightGBM classifier based on SMOTE and elastic net. Anal Biochem 609:113903 https://doi.org/10.1016/j.ab.2020.113903
19. Malebary SJ, Rehman MSU, Khan YD (2019) iCrotoK-PseAAC: Identify lysine crotonylation sites by blending position relative statistical features according to the Chou's 5-step rule. PLoS One 14:e0223993 https://doi.org/10.1371/journal.pone.0223993
20. Lv H, Dao FY, Guan ZX, et al. (2021) Deep-Kcr: accurate detection of lysine crotonylation sites using deep learning method. Brief Bioinform 22 https://doi.org/10.1093/bib/bbaa255
21. Li Z, Fang J, Wang S, et al. (2022) Adapt-Kcr: a novel deep learning framework for accurate prediction of lysine crotonylation sites based on learning embedding features and attention architecture. Brief Bioinform 23 https://doi.org/10.1093/bib/bbac037
22. Qiao Y, Zhu X, Gong H (2022) BERT-Kcr: prediction of lysine crotonylation sites by a transfer learning method with pre-trained BERT models. Bioinformatics 38:648 https://doi.org/10.1093/bioinformatics/btab712
23. Pratyush P, Bahmani S, Pokharel S, et al. (2024) LMCrot: an enhanced protein crotonylation site predictor by

leveraging an interpretable window-level embedding from a transformer-based protein language model. Bioinformatics 40 https://doi.org/10.1093/bioinformatics/btae290
24. Zhang J, Cao H, Gao Z, et al. (2026) Learning the PTM code through a coarse-to-fine mechanism-aware framework. Nat Commun 17 https://doi.org/10.1038/s41467-026-73148-3
25. Dou L, Zhang Z, Xu L, et al. (2022) iKcr_CNN: A novel computational tool for imbalance classification of human nonhistone crotonylation sites based on convolutional neural networks with focal loss. Comput Struct Biotechnol J 20:3268 https://doi.org/10.1016/j.csbj.2022.06.032
26. Wei X, Hu S, Tu J, et al. (2025) iKcr-DRC: prediction of lysine crotonylation sites in proteins based on a novel attention module and DenseNet. Front Genet 16:1574832 https://doi.org/10.3389/fgene.2025.1574832
27. Zhang HQ, Lin XR, Wang YT, et al. (2025) EDS-Kcr: deep supervision based on large language model for identifying protein lysine crotonylation sites across multiple species. Brief Bioinform 26 https://doi.org/10.1093/bib/bbaf249
28. Wei X, Sha Y, Zhao Y, et al. (2021) DeepKcrot: a deep-learning architecture for general and species-specific lysine crotonylation site prediction. Ieee Access 9:49504
29. Khanal J, Kandel J, Tayara H, et al. (2023) CapsNh-Kcr: Capsule network-based prediction of lysine crotonylation sites in human non-histone proteins. Comput Struct Biotechnol J 21:120 https://doi.org/10.1016/j.csbj.2022.11.056
30. UniProt C (2023) UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res 51:D523 https://doi.org/10.1093/nar/gkac1052
31. Sayers EW, Bolton EE, Brister JR, et al. (2023) Database resources of the National Center for Biotechnology Information in 2023. Nucleic Acids Res 51:D29 https://doi.org/10.1093/nar/gkac1032
32. Fu L, Niu B, Zhu Z, et al. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150 https://doi.org/10.1093/bioinformatics/bts565
33. Buschmann L, Bolz SN, El-Hendi F, et al. (2026) Much ado about nothing: modeling amino acid replacement with predicted protein structures. Bioinformatics 42 https://doi.org/10.1093/bioinformatics/btag188
34. Zhang Z, Wayment-Steele HK, Brixi G, et al. (2024) Protein language models learn evolutionary statistics of interacting sequence motifs. Proc Natl Acad Sci U S A 121:e2406285121 https://doi.org/10.1073/pnas.2406285121
35. Yuan L, Hu X, Ma Y, et al. (2022) DLBLS_SS: protein secondary structure prediction using deep learning and broad learning system. RSC Adv 12:33479 https://doi.org/10.1039/d2ra06433b
36. Zhao L, Li J, Zhan W, et al. (2024) Prediction of protein secondary structure by the improved TCN-BiLSTM-MHA model with knowledge distillation. Sci Rep 14:16488 https://doi.org/10.1038/s41598-024-67403-0
37. Woo S, Park J, Lee J-Y, et al. Cbam: Convolutional block attention module. European conference on computer vision: Springer, 2018:3
38. Ganaie MA, Hu M, Malik AK, et al. (2022) Ensemble deep learning: A review. Engineering applications of artificial intelligence 115:105151
39. Mohammed A, Kora R (2023) A comprehensive review on ensemble deep learning: Opportunities and challenges. Journal of King Saud University Computer and Information Sciences 35:757
40. Jia J, Lei R, Qin L, et al. (2023) iEnhancer-DCSV: Predicting enhancers and their strength based on DenseNet and improved convolutional block attention module. Front Genet 14:1132018 https://doi.org/10.3389/fgene.2023.1132018
41. Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21:6 https://doi.org/10.1186/s12864-019-6413-7
42. Wu X, Bartel DP (2017) kpLogo: positional k-mer analysis reveals hidden specificity in biological sequences.

679 Nucleic Acids Res 45:W534 https://doi.org/10.1093/nar/gkx323

680